

# Bayesian Decision Theory

Jason Corso

SUNY at Buffalo

January 2012

# Overview and Plan

- Covering Chapter 2 of DHS.
- Bayesian Decision Theory is a fundamental statistical approach to the problem of pattern classification.
- Quantifies the tradeoffs between various classifications using probability and the costs that accompany such classifications.
- Assumptions:
  - Decision problem is posed in probabilistic terms.
  - All relevant probability values are known.

# Recall the Fish!

- Recall our example from the first lecture on classifying two fish as salmon or sea bass.
- And recall our agreement that any given fish is either a salmon or a sea bass; DHS call this the **state of nature** of the fish.
- Let's define a (probabilistic) variable  $\omega$  that describes the state of nature.

$$\omega = \omega_1 \quad \text{for sea bass} \quad (1)$$

$$\omega = \omega_2 \quad \text{for salmon} \quad (2)$$

- Let's assume this two class case.



Salmon



Sea Bass

# Prior Probability

- The *a priori* or **prior** probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.

# Prior Probability

- The *a priori* or **prior** probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.
- In the fish example, it is the probability that we will see either a salmon or a sea bass next on the conveyor belt.
- Note: The prior may vary depending on the situation.
  - If we get equal numbers of salmon and sea bass in a catch, then the priors are equal, or **uniform**.
  - Depending on the season, we may get more salmon than sea bass, for example.

# Prior Probability

- The *a priori* or **prior** probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.
- In the fish example, it is the probability that we will see either a salmon or a sea bass next on the conveyor belt.
- Note: The prior may vary depending on the situation.
  - If we get equal numbers of salmon and sea bass in a catch, then the priors are equal, or **uniform**.
  - Depending on the season, we may get more salmon than sea bass, for example.
- We write  $P(\omega = \omega_1)$  or just  $P(\omega_1)$  for the prior the next is a sea bass.
- The priors must exhibit exclusivity and exhaustivity. For  $c$  states of nature, or classes:

$$1 = \sum_{i=1}^c P(\omega_i) \quad (3)$$

# Decision Rule From Only Priors

- A **decision rule** prescribes what action to take based on observed input.
- IDEA CHECK: What is a reasonable Decision Rule if
  - the only available information is the prior, and
  - the cost of any incorrect classification is equal?

# Decision Rule From Only Priors

- A **decision rule** prescribes what action to take based on observed input.
- IDEA CHECK: What is a reasonable Decision Rule if
  - the only available information is the prior, and
  - the cost of any incorrect classification is equal?
- Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$ ; otherwise decide  $\omega_2$ .
- What can we say about this decision rule?



# Decision Rule From Only Priors

- A **decision rule** prescribes what action to take based on observed input.
- IDEA CHECK: What is a reasonable Decision Rule if
  - the only available information is the prior, and
  - the cost of any incorrect classification is equal?
- Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$ ; otherwise decide  $\omega_2$ .
- What can we say about this decision rule?
  - Seems reasonable, but it will **always** choose the same fish.
  - If the priors are uniform, this rule will behave poorly.
  - Under the given assumptions, no other rule can do better! (We will see this later on.)

# Features and Feature Spaces

- A **feature** is an observable variable.
- A **feature space** is a set from which we can sample or observe values.
- Examples of features:
  - Length
  - Width
  - Lightness
  - Location of Dorsal Fin
- For simplicity, let's assume that our features are all continuous values.
- Denote a scalar feature as  $x$  and a vector feature as  $\mathbf{x}$ . For a  $d$ -dimensional feature space,  $\mathbf{x} \in \mathbb{R}^d$ .

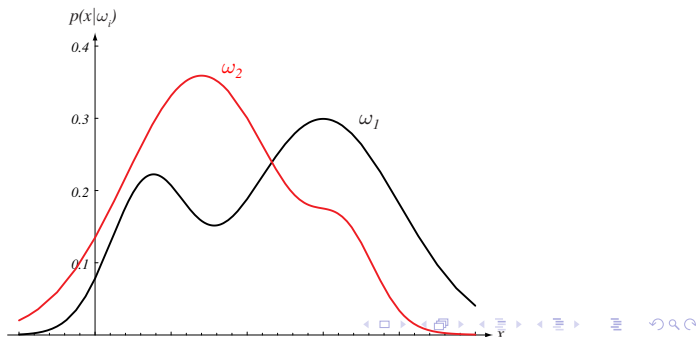
# Class-Conditional Density

## or Likelihood

- The **class-conditional probability density** function is the probability density function for  $\mathbf{x}$ , our feature, given that the state of nature is  $\omega$ :

$$p(\mathbf{x}|\omega) \quad (4)$$

- Here is the hypothetical class-conditional density  $p(x|\omega)$  for lightness values of sea bass and salmon.



# Posterior Probability

## Bayes Formula

- If we know the prior distribution and the class-conditional density, how does this affect our decision rule?
- **Posterior probability** is the probability of a certain state of nature given our observables:  $P(\omega|\mathbf{x})$ .
- Use Bayes Formula:

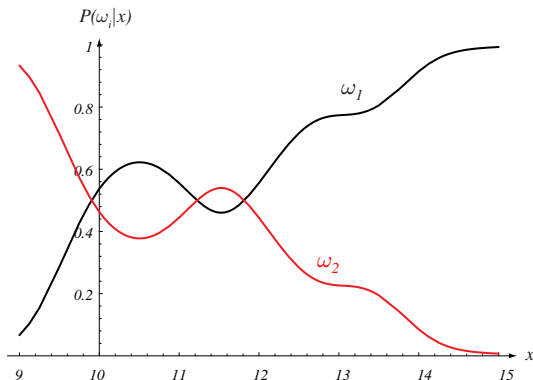
$$P(\omega, \mathbf{x}) = P(\omega|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega)P(\omega) \quad (5)$$

$$P(\omega|\mathbf{x}) = \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})} \quad (6)$$

$$= \frac{p(\mathbf{x}|\omega)P(\omega)}{\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)} \quad (7)$$

# Posterior Probability

- Notice the likelihood and the prior govern the posterior. The  $p(x)$  evidence term is a scale-factor to normalize the density.
- For the case of  $P(\omega_1) = 2/3$  and  $P(\omega_2) = 1/3$  the posterior is



# Probability of Error

- For a given observation  $x$ , we would be inclined to let the posterior govern our decision:

$$\omega^* = \arg \max_i P(\omega_i | \mathbf{x}) \quad (8)$$

- What is our **probability of error**?

# Probability of Error

- For a given observation  $x$ , we would be inclined to let the posterior govern our decision:

$$\omega^* = \arg \max_i P(\omega_i | \mathbf{x}) \quad (8)$$

- What is our **probability of error**?
- For the two class situation, we have

$$P(\text{error} | \mathbf{x}) = \begin{cases} P(\omega_1 | \mathbf{x}) & \text{if we decide } \omega_2 \\ P(\omega_2 | \mathbf{x}) & \text{if we decide } \omega_1 \end{cases} \quad (9)$$

# Probability of Error

- We can minimize the probability of error by following the posterior:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \quad (10)$$



# Probability of Error

- We can minimize the probability of error by following the posterior:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \quad (10)$$

- And, this minimizes the average probability of error too:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (11)$$

(Because the integral will be minimized when we can ensure each  $P(\text{error}|\mathbf{x})$  is as small as possible.)

# Bayes Decision Rule (with Equal Costs)

- Decide  $\omega_1$  if  $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ ; otherwise decide  $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min [P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})] \quad (12)$$

# Bayes Decision Rule (with Equal Costs)

- Decide  $\omega_1$  if  $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ ; otherwise decide  $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min [P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})] \quad (12)$$

- Equivalently, Decide  $\omega_1$  if  $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$
- I.e., the evidence term is not used in decision making.

# Bayes Decision Rule (with Equal Costs)

- Decide  $\omega_1$  if  $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ ; otherwise decide  $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min [P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})] \quad (12)$$

- Equivalently, Decide  $\omega_1$  if  $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$
- I.e., the evidence term is not used in decision making.
- If we have  $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$ , then the decision will rely exclusively on the priors.
- Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.

# Bayes Decision Rule (with Equal Costs)

- Decide  $\omega_1$  if  $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ ; otherwise decide  $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min [P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})] \quad (12)$$

- Equivalently, Decide  $\omega_1$  if  $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$ ; otherwise decide  $\omega_2$
- I.e., the evidence term is not used in decision making.
- If we have  $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$ , then the decision will rely exclusively on the priors.
- Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.
- Take Home Message: **Decision making relies on both the priors and the likelihoods and Bayes Decision Rule combines them to achieve the minimum probability of error.**

# Loss Functions

- A **loss function** states exactly how costly each action is.
- As earlier, we have  $c$  classes  $\{\omega_1, \dots, \omega_c\}$ .
- We also have  $a$  possible actions  $\{\alpha_1, \dots, \alpha_a\}$ .
- The loss function  $\lambda(\alpha_i|\omega_j)$  is the loss incurred for taking action  $\alpha_i$  when the class is  $\omega_j$ .

# Loss Functions

- A **loss function** states exactly how costly each action is.
- As earlier, we have  $c$  classes  $\{\omega_1, \dots, \omega_c\}$ .
- We also have  $a$  possible actions  $\{\alpha_1, \dots, \alpha_a\}$ .
- The loss function  $\lambda(\alpha_i|\omega_j)$  is the loss incurred for taking action  $\alpha_i$  when the class is  $\omega_j$ .
- The **Zero-One Loss Function** is a particularly common one:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, 2, \dots, c \quad (13)$$

It assigns no loss to a correct decision and uniform unit loss to an incorrect decision.

# Expected Loss

a.k.a. Conditional Risk

- We can consider the loss that would be incurred from taking each possible action in our set.
- The **expected loss** or conditional risk is by definition

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \quad (14)$$

- The **zero-one conditional risk** is

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) \quad (15)$$

$$= 1 - P(\omega_i|\mathbf{x}) \quad (16)$$

- Hence, for an observation  $x$ , we can minimize the expected loss by selecting the action that minimizes the conditional risk.



# Expected Loss

a.k.a. Conditional Risk

- We can consider the loss that would be incurred from taking each possible action in our set.
- The **expected loss** or conditional risk is by definition

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \quad (14)$$

- The **zero-one conditional risk** is

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) \quad (15)$$

$$= 1 - P(\omega_i|\mathbf{x}) \quad (16)$$

- Hence, for an observation  $x$ , we can minimize the expected loss by selecting the action that minimizes the conditional risk.
- (Teaser) You guessed it: this is what Bayes Decision Rule does!

# Overall Risk

- Let  $\alpha(x)$  denote a decision rule, a mapping from the input feature space to an action,  $\mathbb{R}^d \mapsto \{\alpha_1, \dots, \alpha_a\}$ .
  - This is what we want to learn.

# Overall Risk

- Let  $\alpha(x)$  denote a decision rule, a mapping from the input feature space to an action,  $\mathbb{R}^d \mapsto \{\alpha_1, \dots, \alpha_a\}$ .
  - This is what we want to learn.
- The **overall risk** is the expected loss associated with a given decision rule.

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (17)$$

Clearly, we want the rule  $\alpha(\cdot)$  that minimizes  $R(\alpha(\mathbf{x})|\mathbf{x})$  for all  $\mathbf{x}$ .

# Bayes Risk

## The Minimum Overall Risk

- Bayes Decision Rule gives us a method for minimizing the overall risk.
- Select the action that minimizes the conditional risk:

$$\alpha^* = \arg \min_{\alpha_i} R(\alpha_i | \mathbf{x}) \quad (18)$$

$$= \arg \min_{\alpha_i} \sum_{j=1}^c \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \quad (19)$$

- The Bayes Risk is the best we can do.

# Two-Category Classification Examples

- Consider two classes and two actions,  $\alpha_1$  when the true class is  $\omega_1$  and  $\alpha_2$  for  $\omega_2$ .
- Writing out the conditional risks gives:

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \quad (20)$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}) . \quad (21)$$

- Fundamental rule is decide  $\omega_1$  if

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x}) . \quad (22)$$

- In terms of posteriors, decide  $\omega_1$  if

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}) . \quad (23)$$

The more likely state of nature is scaled by the differences in loss (which are generally positive).

# Two-Category Classification Examples

- Or, expanding via Bayes Rule, decide  $\omega_1$  if

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2) \quad (24)$$

- Or, assuming  $\lambda_{21} > \lambda_{11}$ , decide  $\omega_1$  if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} \quad (25)$$

- LHS is called the **likelihood ratio**.
- Thus, we can say the Bayes Decision Rule says to decide  $\omega_1$  if the likelihood ratio exceeds a threshold that is independent of the observation  $\mathbf{x}$ .

# Pattern Classifiers Version 1: Discriminant Functions

- **Discriminant Functions** are a useful way of representing pattern classifiers.
- Let's say  $g_i(\mathbf{x})$  is a discriminant function for the  $i$ th class.
- This classifier will assign a class  $\omega_i$  to the feature vector  $\mathbf{x}$  if

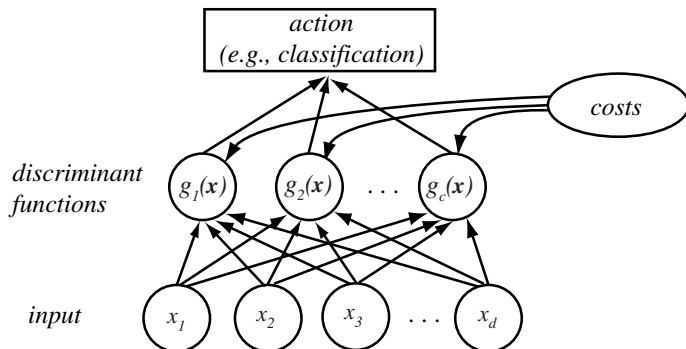
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i, \quad (26)$$

or, equivalently

$$i^* = \arg \max_i g_i(x), \quad \text{decide } \omega_{i^*}.$$

# Discriminants as a Network

- We can view the discriminant classifier as a network (for  $c$  classes and a  $d$ -dimensional input vector).





# Bayes Discriminants

## Minimum Conditional Risk Discriminant

- General case with risks

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) \quad (27)$$

$$= -\sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \quad (28)$$

- Can we prove that this is correct?

# Bayes Discriminants

## Minimum Conditional Risk Discriminant

- General case with risks

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) \quad (27)$$

$$= - \sum_{j=1}^c \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \quad (28)$$

- Can we prove that this is correct?
- **Yes!** The minimum conditional risk corresponds to the maximum discriminant.

# Minimum Error-Rate Discriminant

- In the case of zero-one loss function, the Bayes Discriminant can be further simplified:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \ . \quad (29)$$

# Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?

# Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?
- **No!**
- Multiply by some positive constant.
- Shift them by some additive constant.

# Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?
- **No!**
- Multiply by some positive constant.
- Shift them by some additive constant.
- For monotonically increasing function  $f(\cdot)$ , we can replace each  $g_i(\mathbf{x})$  by  $f(g_i(\mathbf{x}))$  without affecting our classification accuracy.
  - These can help for ease of understanding or computability.
  - The following all yield the same exact classification results for minimum-error-rate classification.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_j p(\mathbf{x}|\omega_j)P(\omega_j)} \quad (30)$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \quad (31)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \quad (32)$$

# Visualizing Discriminants

## Decision Regions

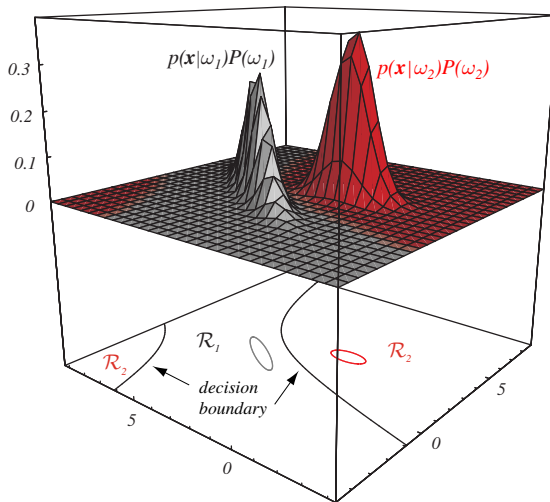
- The effect of any decision rule is to divide the feature space into decision regions.
- Denote a decision region  $\mathcal{R}_i$  for  $\omega_i$ .
- One not necessarily connected region is created for each category and assignments is according to:

$$\text{If } g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i, \text{ then } \mathbf{x} \text{ is in } \mathcal{R}_i . \quad (33)$$

- **Decision boundaries** separate the regions; they are ties among the discriminant functions.

# Visualizing Discriminants

## Decision Regions





# Two-Category Discriminants

## Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad . \quad (34)$$

- What is a suitable decision rule?

# Two-Category Discriminants

## Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad . \quad (34)$$

- The following simple rule is then used:

$$\text{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0; \text{ otherwise decide } \omega_2. \quad (35)$$

# Two-Category Discriminants

## Dichotomizers

- In the two-category case, one considers single discriminant

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \quad . \quad (34)$$

- The following simple rule is then used:

$$\text{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0; \text{ otherwise decide } \omega_2. \quad (35)$$

- Various manipulations of the discriminant:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \quad (36)$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \quad (37)$$

# Background on the Normal Density

- This next section is a slight digression to introduce the Normal Density (most of you will have had this already).
- The Normal density is very well studied.
- It easy to work with analytically.
- Often in PR, an appropriate model seems to be a single typical value corrupted by continuous-valued, random noise.
- Central Limit Theorem (Second Fundamental Theorem of Probability).
  - The distribution of the sum of  $n$  random variables approaches the normal distribution when  $n$  is large.
  - E.g., <http://www.stattucino.com/berrie/dsl/Galton.html>

# Expectation

- Recall the definition of expected value of any scalar function  $f(x)$  in the continuous  $p(x)$  and discrete  $P(x)$  cases

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx \quad (38)$$

$$\mathcal{E}[f(x)] = \sum_x f(x)P(x) \quad (39)$$

where we have a set  $\mathcal{D}$  over which the discrete expectation is computed.

# Univariate Normal Density

- Continuous univariate normal, or **Gaussian**, density:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] . \quad (40)$$

- The **mean** is the expected value of  $x$  is

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)dx . \quad (41)$$

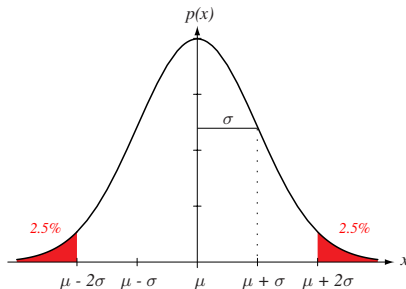
- The **variance** is the expected squared deviation

$$\sigma^2 \equiv \mathcal{E}[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx . \quad (42)$$

# Univariate Normal Density

## Sufficient Statistics

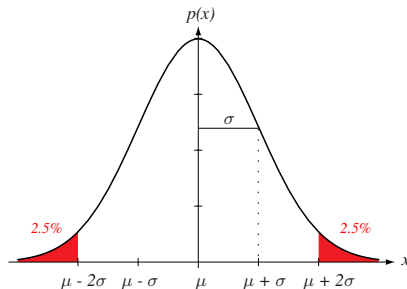
- Samples from the normal density tend to cluster around the mean and be spread-out based on the variance.



# Univariate Normal Density

## Sufficient Statistics

- Samples from the normal density tend to cluster around the mean and be spread-out based on the variance.



- The normal density is completely specified by the mean and the variance. These two are its **sufficient statistics**.
- We thus abbreviate the equation for the normal density as

$$p(x) \sim N(\mu, \sigma^2) \quad (43)$$



# Entropy

- **Entropy** is the uncertainty in the random samples from a distribution.

$$H(p(x)) = - \int p(x) \ln p(x) dx \quad (44)$$

- The normal density has the maximum entropy for all distributions have a given mean and variance.
- What is the entropy of the uniform distribution?

# Entropy

- **Entropy** is the uncertainty in the random samples from a distribution.

$$H(p(x)) = - \int p(x) \ln p(x) dx \quad (44)$$

- The normal density has the maximum entropy for all distributions have a given mean and variance.
- What is the entropy of the uniform distribution?
- The uniform distribution has maximum entropy (on a given interval).

# Multivariate Normal Density

And a test to see if your Linear Algebra is up to snuff.

- The multivariate Gaussian in  $d$  dimensions is written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right] . \quad (45)$$

- Again, we abbreviate this as  $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \Sigma)$ .
- The sufficient statistics in  $d$ -dimensions:

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad (46)$$

$$\Sigma \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x} \quad (47)$$

# The Covariance Matrix

$$\Sigma \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements  $\sigma_{ii}$  are the variances of the respective coordinate  $x_i$ .
- The off-diagonal elements  $\sigma_{ij}$  are the covariances of  $x_i$  and  $x_j$ .
- What does a  $\sigma_{ij} = 0$  imply?

# The Covariance Matrix

$$\Sigma \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements  $\sigma_{ii}$  are the variances of the respective coordinate  $x_i$ .
- The off-diagonal elements  $\sigma_{ij}$  are the covariances of  $x_i$  and  $x_j$ .
- What does a  $\sigma_{ij} = 0$  imply?
- That coordinates  $x_i$  and  $x_j$  are statistically independent.

# The Covariance Matrix

$$\Sigma \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements  $\sigma_{ii}$  are the variances of the respective coordinate  $x_i$ .
- The off-diagonal elements  $\sigma_{ij}$  are the covariances of  $x_i$  and  $x_j$ .
- What does a  $\sigma_{ij} = 0$  imply?
  - That coordinates  $x_i$  and  $x_j$  are statistically independent.
- What does  $\Sigma$  reduce to if all off-diagonals are 0?

# The Covariance Matrix

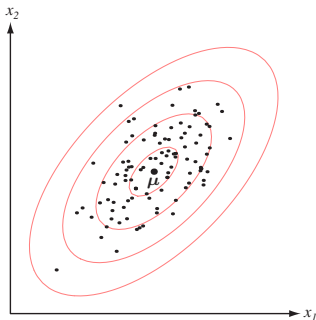
$$\Sigma \equiv \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}$$

- Symmetric.
- Positive semi-definite (but DHS only considers positive definite so that the determinant is strictly positive).
- The diagonal elements  $\sigma_{ii}$  are the variances of the respective coordinate  $x_i$ .
- The off-diagonal elements  $\sigma_{ij}$  are the covariances of  $x_i$  and  $x_j$ .
- What does a  $\sigma_{ij} = 0$  imply?
  - That coordinates  $x_i$  and  $x_j$  are statistically independent.
- What does  $\Sigma$  reduce to if all off-diagonals are 0?
- The product of the  $d$  univariate densities.

# Mahalanobis Distance

- The shape of the density is determined by the covariance  $\Sigma$ .
- Specifically, the eigenvectors of  $\Sigma$  give the principal axes of the hyperellipsoids and the eigenvalues determine the lengths of these axes.
- The loci of points of constant density are hyperellipsoids with constant **Mahalanobis distance**:

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (48)$$



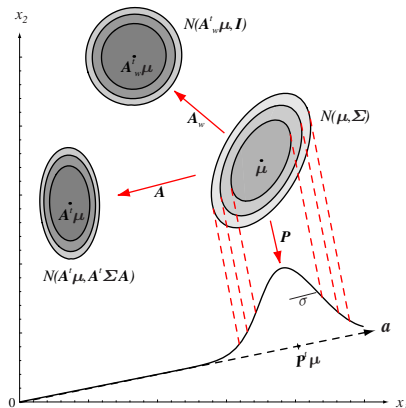


# Linear Combinations of Normals

- Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed.
- For  $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\mathbf{A}$ , a  $d$ -by- $k$  matrix, define  $\mathbf{y} = \mathbf{A}^T \mathbf{x}$ . Then:

$$p(\mathbf{y}) \sim N(\mathbf{A}^T \boldsymbol{\mu}, \mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A}) \quad (49)$$

- With the covariance matrix, we can calculate the dispersion of the data in any direction or in any subspace.



# General Discriminant for Normal Densities

- Recall the minimum error rate discriminant,  
 $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i).$
- If we assume normal densities, i.e., if  $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ , then the general discriminant is of the form

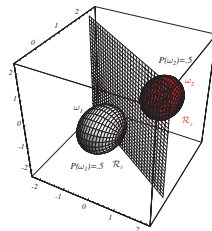
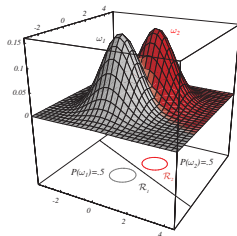
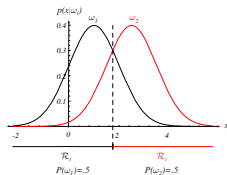
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (50)$$

# Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume  $\Sigma_i = \sigma^2 \mathbf{I}$ ?

# Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume  $\Sigma_i = \sigma^2 \mathbf{I}$ ?
- They are hyperplanes.



- Let's see why...

# Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

- The discriminant functions take on a simple form:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i) \quad (51)$$

- Think of this discriminant as a combination of two things
  - 1 The distance of the sample to the mean vector (for each  $i$ ).
  - 2 A normalization by the variance and offset by the prior.

# Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

- But, we don't need to actually compute the distances.
- Expanding the quadratic form  $(\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu})$  yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} \left[ \mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\mu}_i^\top \mathbf{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right] + \ln P(\omega_i) . \quad (52)$$

- The quadratic term  $\mathbf{x}^\top \mathbf{x}$  is the same for all  $i$  and can thus be ignored.
- This yields the equivalent **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x} + w_{i0} \quad (53)$$

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \quad (54)$$

$$w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i + \ln P(\omega_i) \quad (55)$$

- $w_{i0}$  is called the **bias**.

# Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

## Decision Boundary Equation

- The decision surfaces for a linear discriminant classifiers are hyperplanes defined by the linear equations  $g_i(\mathbf{x}) = g_j(\mathbf{x})$ .
- The equation can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad (56)$$

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \quad (57)$$

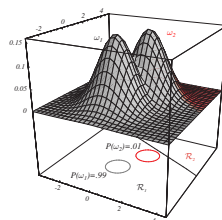
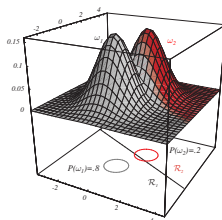
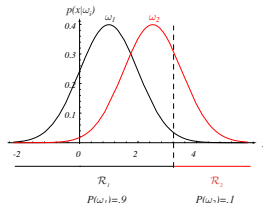
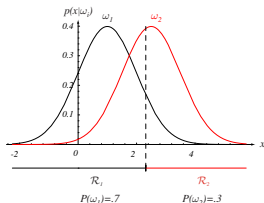
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \quad (58)$$

- These equations define a hyperplane through point  $x_0$  with a normal vector  $\mathbf{w}$ .

# Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

## Decision Boundary Equation

- The decision boundary changes with the prior.





# General Case: Arbitrary $\Sigma_i$

- The discriminant functions are quadratic (the only term we can drop is the  $\ln 2\pi$  term):

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (59)$$

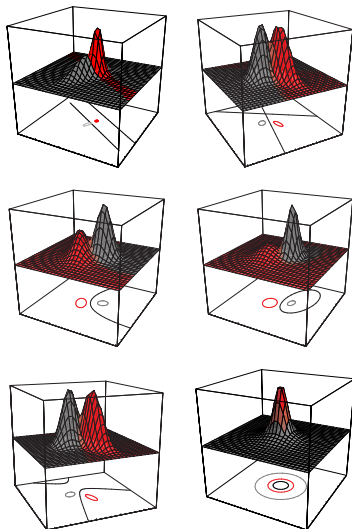
$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1} \quad (60)$$

$$\mathbf{w}_i = \Sigma_i^{-1} \boldsymbol{\mu}_i \quad (61)$$

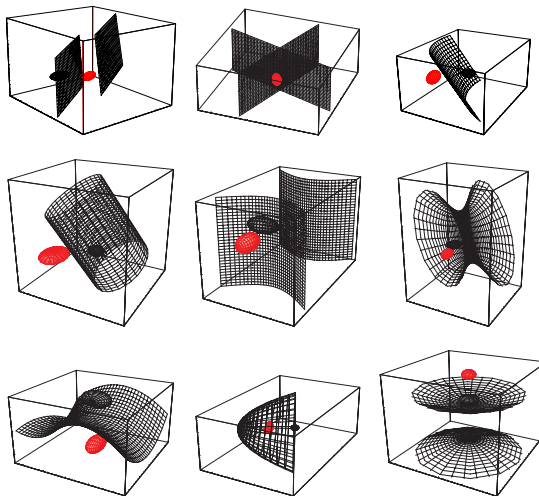
$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \quad (62)$$

- The decision surface between two categories are **hyperquadrics**.

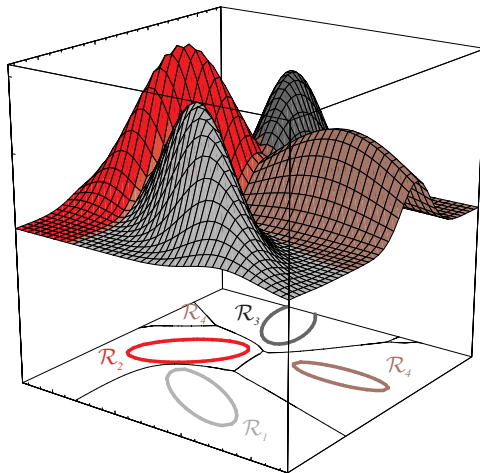
# General Case: Arbitrary $\Sigma_i$



# General Case: Arbitrary $\Sigma_i$



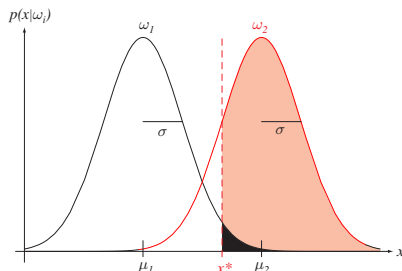
# General Case for Multiple Categories



**Quite A Complicated Decision Surface!**

# Signal Detection Theory

- A fundamental way of analyzing a classifier.
- Consider the following experimental setup:



- Suppose we are interested in detecting a single pulse.
- We can read an internal signal  $x$ .
- The signal is distributed about mean  $\mu_2$  when an external signal is present and around mean  $\mu_1$  when no external signal is present.
- Assume the distributions have the same variances,  $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$ .

# Signal Detection Theory

- The detector uses  $x^*$  to decide if the external signal is present.
- **Discriminability** characterizes how difficult it will be to decide if the external signal is present without knowing  $x^*$ .

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \quad (63)$$

- Even if we do not know  $\mu_1$ ,  $\mu_2$ ,  $\sigma$ , or  $x^*$ , we can find  $d'$  by using a **receiver operating characteristic** or ROC curve, as long as we know the state of nature for some experiments

# Receiver Operating Characteristics

## Definitions

- A **Hit** is the probability that the internal signal is above  $x^*$  given that the external signal is present

$$P(x > x^* | x \in \omega_2) \quad (64)$$

# Receiver Operating Characteristics

## Definitions

- A **Hit** is the probability that the internal signal is above  $x^*$  given that the external signal is present

$$P(x > x^* | x \in \omega_2) \quad (64)$$

- A **Correct Rejection** is the probability that the internal signal is below  $x^*$  given that the external signal is not present.

$$P(x < x^* | x \in \omega_1) \quad (65)$$



# Receiver Operating Characteristics

## Definitions

- A **Hit** is the probability that the internal signal is above  $x^*$  given that the external signal is present

$$P(x > x^* | x \in \omega_2) \quad (64)$$

- A **Correct Rejection** is the probability that the internal signal is below  $x^*$  given that the external signal is not present.

$$P(x < x^* | x \in \omega_1) \quad (65)$$

- A **False Alarm** is the probability that the internal signal is above  $x^*$  despite there being no external signal present.

$$P(x > x^* | x \in \omega_1) \quad (66)$$

# Receiver Operating Characteristics

## Definitions

- A **Hit** is the probability that the internal signal is above  $x^*$  given that the external signal is present

$$P(x > x^* | x \in \omega_2) \quad (64)$$

- A **Correct Rejection** is the probability that the internal signal is below  $x^*$  given that the external signal is not present.

$$P(x < x^* | x \in \omega_1) \quad (65)$$

- A **False Alarm** is the probability that the internal signal is above  $x^*$  despite there being no external signal present.

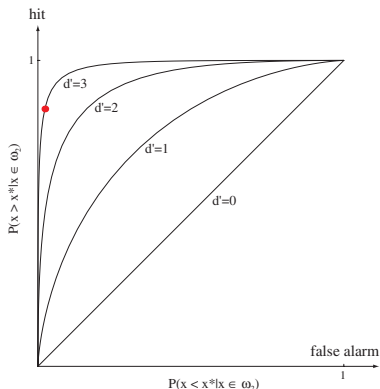
$$P(x > x^* | x \in \omega_1) \quad (66)$$

- A **Miss** is the probability that the internal signal is below  $x^*$  given that the external signal is present.

$$P(x < x^* | x \in \omega_2) \quad (67)$$

# Receiver Operating Characteristics

- We can experimentally determine the rates, in particular the Hit-Rate and the False-Alarm-Rate.
- Basic idea is to assume our densities are fixed (reasonable) but vary our threshold  $x^*$ , which will thus change the rates.
- The receiver operating characteristic plots the hit rate against the false alarm rate.
- What shape curve do we want?



# Missing Features

- Suppose we have built a classifier on multiple features, for example the lightness and width.
- What do we do if one of the features is not measurable for a particular case? For example the lightness can be measured but the width cannot because of occlusion.

# Missing Features

- Suppose we have built a classifier on multiple features, for example the lightness and width.
- What do we do if one of the features is not measurable for a particular case? For example the lightness can be measured but the width cannot because of occlusion.
- **Marginalize!**
- Let  $\mathbf{x}$  be our full feature feature and  $\mathbf{x}_g$  be the subset that are measurable (or good) and let  $\mathbf{x}_b$  be the subset that are missing (or bad/noisy).
- We seek an estimate of the posterior given **just the good features**  $\mathbf{x}_g$ .

# Missing Features

$$P(\omega_i | \mathbf{x}_g) = \frac{p(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} \quad (68)$$

$$= \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}{p(\mathbf{x}_g)} \quad (69)$$

$$= \frac{\int p(\omega_i | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}_b}{p(\mathbf{x}_g)} \quad (70)$$

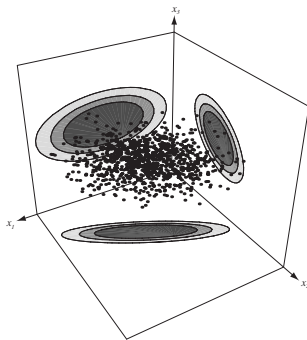
$$= \frac{\int g_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}_b}{\int p(\mathbf{x}) d\mathbf{x}_b} \quad (71)$$

- We will cover the Expectation-Maximization algorithm later.
- This is normally quite expensive to evaluate unless the densities are special (like Gaussians).

# Statistical Independence

- Two variables  $x_i$  and  $x_j$  are independent if

$$p(x_i, x_j) = p(x_i)p(x_j) \quad (72)$$

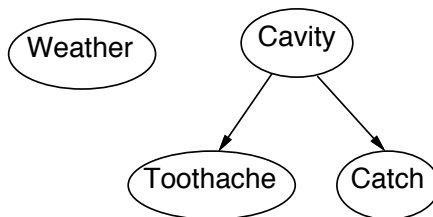


**FIGURE 2.23.** A three-dimensional distribution which obeys  $p(x_1, x_3) = p(x_1)p(x_3)$ ; thus here  $x_1$  and  $x_3$  are statistically independent but the other feature pairs are not. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# Simple Example of Conditional Independence

From Russell and Norvig

- Consider a simple example consisting of four variables: the weather, the presence of a cavity, the presence of a toothache, and the presence of other mouth-related variables such as dry mouth.
- The weather is clearly independent of the other three variables.
- And the toothache and catch are conditionally independent given the cavity (one has no effect on the other given the information about the cavity).





# Naïve Bayes Rule

- If we assume that all of our individual features  $x_i, i = 1, \dots, d$  are conditionally independent given the class, then we have

$$p(\omega_k | \mathbf{x}) \propto \prod_{i=1}^d p(x_i | \omega_k) \quad (73)$$

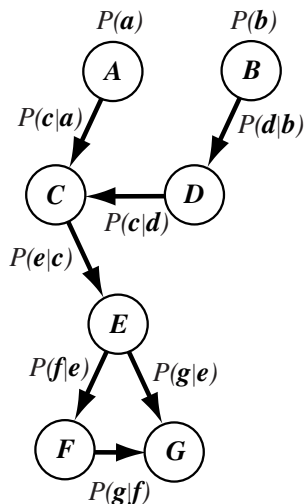
- Circumvents issues of dimensionality.
- Performs with surprising accuracy even in cases violating the underlying independence assumption.

# An Early Graphical Model

- We represent these statistical dependencies graphically.
- Bayesian Belief Networks, or Bayes Nets, are **directed acyclic graphs**.
- Each link is directional.
- No loops.
- The Bayes Net factorizes the distribution into independent parts (making for more easily learned and computed terms).

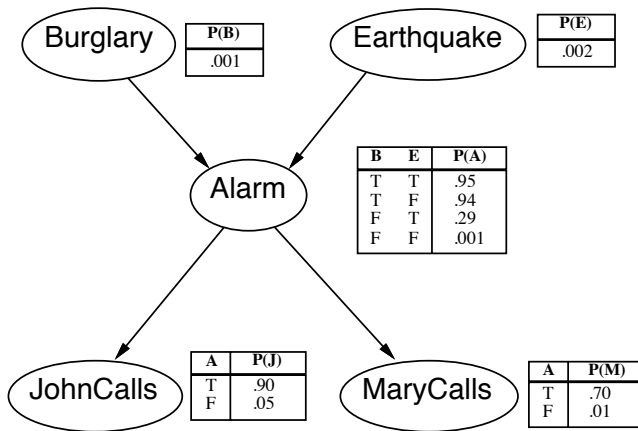
# Bayes Nets Components

- Each **node** represents one variable (assume discrete for simplicity).
- A **link** joining two nodes is directional and it represents **conditional probabilities**.
- The intuitive meaning of a link is that the source has a direct influence on the sink.
- Since we typically work with discrete distributions, we evaluate the conditional probability at each node given its parents and store it in a lookup table called a **conditional probability table**.



# A More Complex Example

From Russell and Norvig



- Key: given knowledge of the values of some nodes in the network, we can apply Bayesian inference to determine the maximum posterior values of the unknown variables!

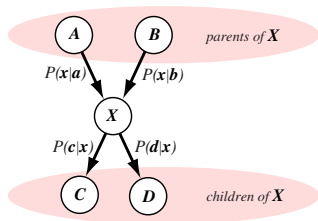
# Full Joint Distribution on a Bayes Net

- Consider a Bayes network with  $n$  variables  $x_1, \dots, x_n$ .
- Denote the parents of a node  $x_i$  as  $\mathcal{P}(x_i)$ .
- Then, we can decompose the joint distribution into the product of conditionals

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \mathcal{P}(x_i)) \quad (74)$$

# Belief at a Single Node

- What is the distribution at a single node, given the rest of the network and the evidence  $\mathbf{e}$ ?
- Parents** of  $\mathbf{X}$ , the set  $\mathcal{P}$  are the nodes on which  $\mathbf{X}$  is conditioned.
- Children** of  $\mathbf{X}$ , the set  $\mathcal{C}$  are the nodes conditioned on  $\mathbf{X}$ .
- Use the Bayes Rule, for the case on the right:



$$P(a, b, x, c, d) = P(a, b, x|c, d)P(c, d) \quad (75)$$

$$= P(a, b|x)P(x|c, d)P(c, d) \quad (76)$$

or more generally,

$$P(\mathcal{C}(x), x, \mathcal{P}(x)|\mathbf{e}) = P(\mathcal{C}(x)|x, \mathbf{e})P(x|\mathcal{P}(x), \mathbf{e})P(\mathcal{P}(x)|, \mathbf{e}) \quad (77)$$