Bayesian Decision Theory
THE GENERAL CASE

- **Finite set of *c* states of nature**
  - $\omega_1 \ldots \omega_c$
  - priors $P(\omega_1) \ldots P(\omega_c)$
- **Measurement is a feature vector**
  - $\mathbf{x} \in \Re^k$
  - *k*-dimensional Euclidean feature space
  - likelihoods $p(\mathbf{x}|\omega_1) \ldots p(\mathbf{x}|\omega_c)$

Bayesian Decision Theory
THE GENERAL CASE

- ***a* different actions are possible**
  - $\alpha_1 \ldots \alpha_a$
  - Loss functions $\lambda(\alpha_1|\omega_1) \ldots \lambda(\alpha_a| \omega_c)$

Bayesian Decision Theory
THE GENERAL CASE

- **Bayes's formula again gives:**

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{p(\mathbf{x})}$$

- **Evidence now is:**

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j)P(\omega_j)$$

Bayesian Decision Theory
THE GENERAL CASE

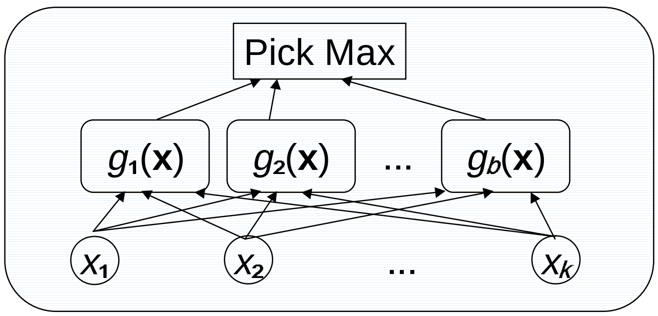
- **Risk (expected loss) defined as:**

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

- **Decision algorithm:**
  - given **x**
  - choose α_i for which $R(\alpha_i \mid \mathbf{x})$ is minimum

Building classifiers

- **Bayesian decision theory leads to the following picture of a classifier:**

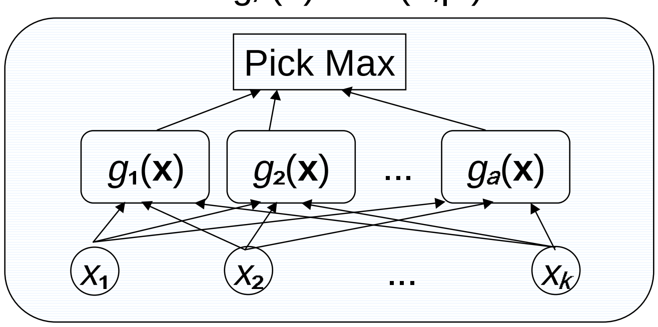


- **g_i are called discriminant functions**

Sample Discriminant Functions

- **Risk:**

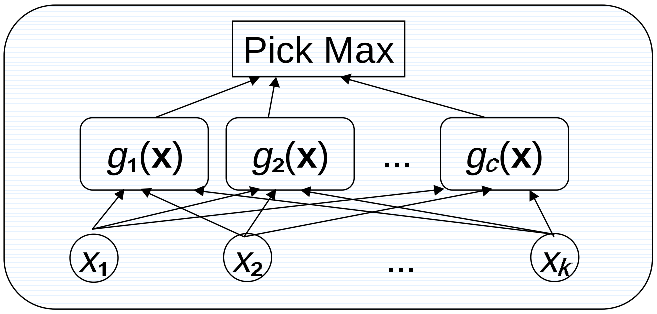
$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$$



Sample Discriminant Functions

- **Likelihood:**

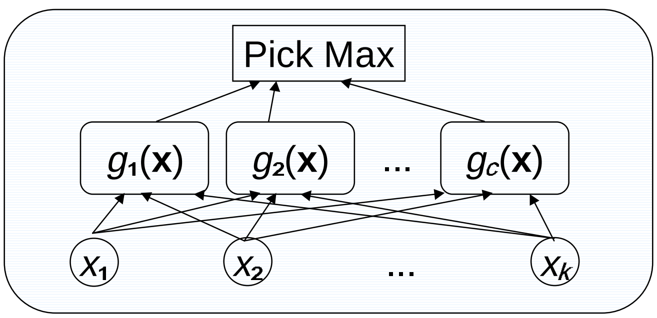
$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$$



Sample Discriminant Functions

- **Non-normalized likelihood:**

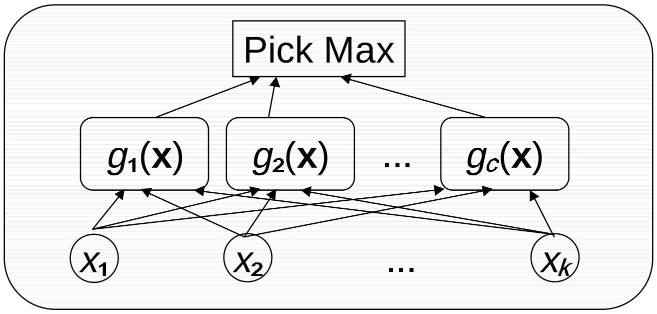
$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$



Sample Discriminant Functions

- **Non-normalized log likelihood:**

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



Implementing discriminant functions

- **Encodes and exploits assumptions about the distributions of measurements**
- **Important special case: normally distributed measurements**

Understanding distributions
CONTINUOUS SCALAR CASE

- **Expected value of a scalar function**

  $E[f(x)] \equiv \int f(x)\, p(x)\, dx$

- **Mean (expected value of *x*)**

  $\mu = E[x] = \int x\, p(x)\, dx$

- **Variance (expected squared deviation)**

  $\sigma^2 = E[(x - \mu)^2] = \int (x - \mu)^2\, p(x)\, dx$

- **Entropy (negative expected log density)**

  $H(p(x)) = - E[\ln p(x)] = - \int p(x) \ln p(x)\, dx$

Univariate Normal Density

- **Defined as**

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

- **Shorthand:**

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

Simple Classification Example

- **Binary decision from one measurement**
  - $P(\omega_1)$, $P(\omega_2)$
  - $p(x|\omega_1) \sim N(\mu_1,\sigma_1^2)$, $p(x|\omega_2) \sim N(\mu_2,\sigma_2^2)$
  - $g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$

Working out the details…

- **Calculate** $g_i(x)$ **as:**

$$g_i(x) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]\right) + \ln P(\omega_i)$$

$$= -\frac{1}{2}\ln 2\pi - \ln\sigma_i - \frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2 + \ln P(\omega_i)$$

$$\approx -\frac{(x-\mu_i)^2}{2\sigma_i^2} - \ln\sigma_i + \ln P(\omega_i)$$

Decide ω_1 if $g_1(x) > g_2(x)$

- **In other words if:**

$$-\frac{(x-\mu_1)^2}{2\sigma_1^{\ 2}} - \ln\left(\frac{\sigma_1}{P(\omega_1)}\right) > -\frac{(x-\mu_2)^2}{2\sigma_2^{\ 2}} - \ln\left(\frac{\sigma_2}{P(\omega_2)}\right)$$

$$\frac{(x-\mu_1)^2}{2\sigma_1^{\ 2}} < \frac{(x-\mu_2)^2}{2\sigma_2^{\ 2}} + \ln\left(\frac{\sigma_2 P(\omega_1)}{\sigma_1 P(\omega_2)}\right)$$

$$\frac{(x-\mu_1)^2}{2\sigma_1^{\ 2}} < \frac{(x-\mu_2)^2}{2\sigma_2^{\ 2}} + t$$

Understanding distributions
MULTIVARIATE CASE

- **Expected value of a scalar function**
  - integrate over the whole feature space:

    $E[f(\mathbf{x})] \equiv \int f(\mathbf{x})\ p(\mathbf{x})\ d\mathbf{x}$
- **For vectors, matrices acts componentwise**
  - Mean (expected value of **x**)

    $\mu = E[\mathbf{x}] = \int \mathbf{x}\ p(\mathbf{x})\ d\mathbf{x}$

    $\mu_i = E[x_i] = \int x_i\ p(\mathbf{x})\ d\mathbf{x}$

Understanding distributions
MULTIVARIATE CASE, CTD

- **Covariance (deviation+correlation):**

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

$$\sigma_{ij} \equiv E\left[(x_i-\mu_i)(x_j-\mu_j)\right] = \int (x_i-\mu_i)(x_j-\mu_j)p(\mathbf{x})d\mathbf{x}$$

Multivariate Normal Density

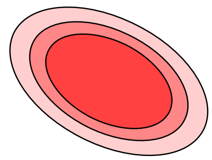
- **Defined as**

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

- **Shorthand:**

$$p(\mathbf{x}) \sim N(\mu, \Sigma)$$

A Quick Visualization



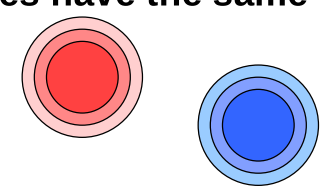
- **Probability density falls off in hyperellipsoids of constant Mahalonobis distance**

$$r^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$$

- **Covariance determines rotation and shape**

Sample multivariate classification case

- **Features are statistically independent, across categories**
- **Features have the same variance σ^2**



- **Geometric intuition: categories determine equal-size hyperspherical clusters**

Formal description

- **Conditional pdf is normal for each class**

$$p(\mathbf{x} \mid \omega_j) \sim N(\mu_j, \Sigma)$$

  - mean vector for each class: μ_j
  - covariance matrix $\Sigma = \sigma^2\mathbf{I}$
  - determinant $|\Sigma| = \sigma^{2k}$
  - inverse $\Sigma^{-1} = (1/\sigma^2)\mathbf{I}$

Working it through

- **By algebra as in univariate case we get:**

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

- **But it's not necessary to compute distances:**

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\mu^T\mathbf{x} + \mu^T\mu\right] + \ln P(\omega_i)$$

$\mathbf{x}^T\mathbf{x}$ is the same for all i

Linear Discriminant Functions

- **So equivalently:**

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

- **For weight vector**

$$\mathbf{w}_i^{\mathrm{T}} = \frac{1}{\sigma^2}\mu_i$$

- **and threshold or bias**

$$w_{i0} = \frac{-1}{2\sigma^2}\mu^{\mathrm{T}}\mu + \ln P(\omega_i)$$

The behavior of classifiers

- **Decision rule divides feature space into decision regions**
  - If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all j, then $\mathbf{x}$ is in region i
  - Regions separated by decision boundaries (where largest discriminant functions tie)

The behavior of linear classifiers

- **Decision surfaces are pieces of hyperplanes** g_i **(x)** = g_j **(x)**
  – Orthogonal to the line between the means
  – Shifted from halfway by variance and priors
- **Explicitly:** $\mathbf{w}^{T}(\mathbf{x}-\mathbf{x}_0)=0$

$$\mathbf{w}=\mu_i-\mu_j$$

$$\mathbf{x}_0=\frac{1}{2}(\mu_i+\mu_j)-\frac{\sigma^2}{\|\mu_i-\mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i-\mu_j)$$