COSC 522 – Machine Learning

Discriminant Functions

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
https://www.eecs.utk.edu/people/hairong-qi/
Email: hqi@utk.edu

Course Website: http://web.eecs.utk.edu/~hqi/cosc522/

Recap from Previous Lecture

- Definition of supervised learning (vs. unsupervised learning)
- The difference between the training set and the test set
- The difference between classification and regression
- Definition of “features”, “samples”, and “dimension”
- From histogram to probability density function (pdf)
- In Bayes’ Formula, what is conditional pdf? Prior probability? Posterior probability?
- What does the normalization factor (or evidence) do?
- What is Baysian decision rule? or MPP?
- What are decision regions?
- How to calculate conditional probability of error and overall probability of error?
- What are cost function (or objective function) and optimization method used in MPP?

terminologies

the Formula

decision rule

Recap

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

Maximum Posterior Probability

For a given x, if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then x belongs to class 1, otherwise, 2.

Overall probability of error

$$P(error) = \int_{\Re_1} P(\omega_2 \mid x) p(x) dx + \int_{\Re_2} P(\omega_1 \mid x) p(x) dx$$

Decision Rule → Decision Region →
Conditional Probability of Error → Overall Probability of Error

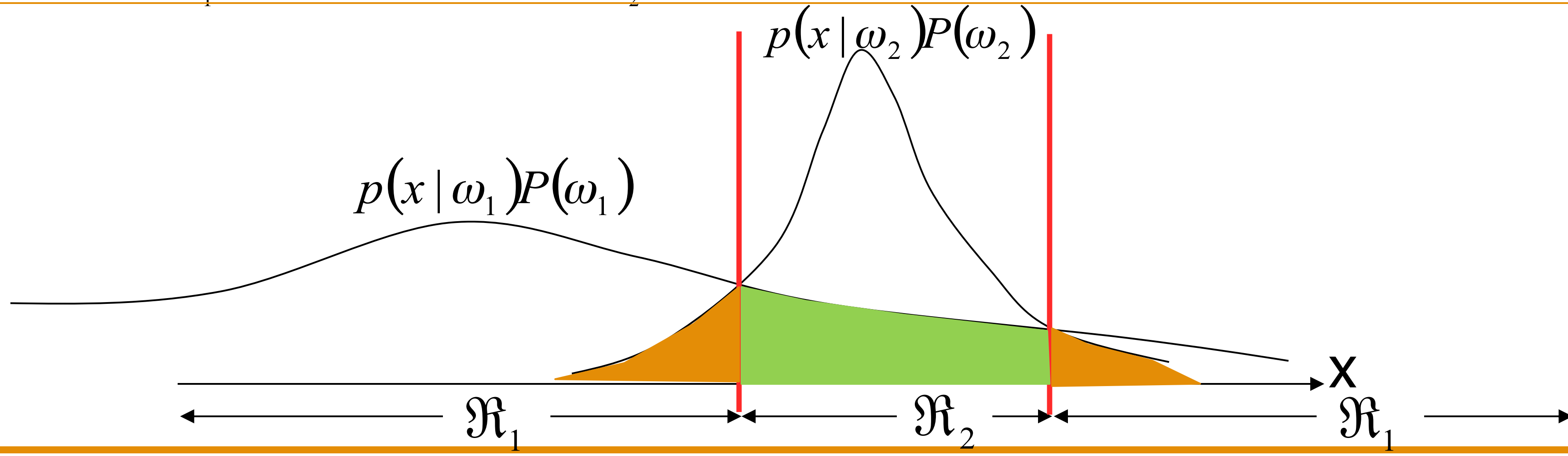
AICIP RESEARCH

$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{if we decide } \omega_2 \\ P(\omega_2 \mid x) & \text{if we decide } \omega_1 \end{cases} = \min\left[P(\omega_1 \mid x), P(\omega_2 \mid x)\right]$$

Unconditional risk, unconditional probability of error

$$P(error) = \int_{-\infty}^{\infty} P(error, x)\,dx = \int_{-\infty}^{\infty} P(error \mid x)\,p(x)\,dx$$

$$P(error) = \int_{\Re_1} P(\omega_2 \mid x)\,p(x)\,dx + \int_{\Re_2} P(\omega_1 \mid x)\,p(x)\,dx = P(error|\omega_2) + P(error|\omega_1)$$



Questions

AICIP RESEARCH

- What is a discriminant function?
- What is a multivariate Gaussian (or normal density function)?
- What is the covariance matrix and what is its dimension?
- What would the covariance matrix look like if the features are independent from each other?
- What would the covariance matrix look like if the features are independent from each other AND have the same spread in each dimension?
- What is minimum (Euclidean) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Euclidean) distance classifier?
- What is minimum (Mahalanobis) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Mahalanobis) distance classifier?
- What does the decision boundary look like for a quadratic classifier?
- What are the cost functions for the discriminant functions? And what is the optimization method used to find the best solution?

Multi-variate Gaussian

Linear and Quadratic Machines and their assumptions

Discrimimant Function

AICIP RESEARCH

- One way to represent pattern classifier- use discriminant functions $g_i(x)$

$$g_i(x) = P(\omega_i|x)$$

$$g_i(x) = p(x|\omega_i)P(\omega_i)$$

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

The classifier will assign a feature vector x to class ω_i if

$$g_i(x) > g_j(x)$$

- For two-class cases,

$$g(x) = g_1(x) - g_2(x) = P(\omega_1 | x) - P(\omega_2 | x)$$

Multivariate Normal Density

AICIP RESEARCH

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

$\vec{x}$: d - component column vector

$\vec{\mu}$: d - component mean vector

Σ : d - by - d covariance matrix

$|\Sigma|$: determinant

Σ^{-1} : inverse

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}, \vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1{}^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d{}^2 \end{bmatrix}$$

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

Estimating Normal Densities

- Calculate μ, Σ

$$\vec{\mu}_i = \begin{bmatrix} \mu_{i1} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{k1} \\ \vdots \\ \mu_{id} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{kd} \end{bmatrix}$$

$$\Sigma_i = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix} = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(\vec{x}_k - \vec{\mu}_i\right)\left(\vec{x}_k - \vec{\mu}_i\right)^T$$

Covariance

For d sets of variates denoted $\{x_1\}, \cdots, \{x_p\}, \cdots, \{x_q\}, \cdots, \{x_d\}$,
the covariance $\sigma_{pq} = \operatorname{cov}(x_p, x_q)$ of x_p and x_q is defined by

$$
\begin{aligned}
\operatorname{cov}(x_p, x_q) &= E\left[(x_p - \mu_p)(x_q - \mu_q)\right] \\
&= E[x_p x_q] - E[x_p \mu_q] - E[\mu_p x_q] + E[\mu_p \mu_q] \\
&= E[x_p x_q] - \mu_q E[x_p] - \mu_p E[x_q] + \mu_p \mu_q \\
&= E[x_p x_q] - \mu_q \mu_p - \mu_p \mu_q + \mu_p \mu_q \\
&= E[x_p x_q] - \mu_q \mu_p
\end{aligned}
$$

When $p = q, \sigma_{pp} = \operatorname{cov}(x_p, x_p) = E[x_p x_p] - \mu_p \mu_p$

$$
\begin{aligned}
&= E\left[x_p{}^2\right] - \left(E[x_p]\right)^2 \\
&= \sigma_p^2
\end{aligned}
$$

Discriminant Function for Normal Density

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

$$\begin{aligned} g_i(\vec{x}) &= \ln p(\vec{x} \mid \omega_i) + \ln P(\omega_i) \\ &= -\frac{1}{2} (\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) - \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i) \\ &= -\frac{1}{2} (\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i) \end{aligned}$$

Questions

AICIP RESEARCH

- What is a discriminant function?
- What is a multivariate Gaussian (or normal density function)?
- What is the covariance matrix and what is its dimension?
- What would the covariance matrix look like if the features are independent from each other?
- What would the covariance matrix look like if the features are independent from each other AND have the same spread in each dimension?
- What is minimum (Euclidean) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Euclidean) distance classifier?
- What is minimum (Mahalanobis) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Mahalanobis) distance classifier?
- What does the decision boundary look like for a quadratic classifier?
- What are the cost functions for the discriminant functions? And what is the optimization method used to find the best solution?

Multi-variate Gaussian

Linear and Quadratic Machines and their assumptions

Case 1: $\Sigma_i=\sigma^2 I$

- The features are statistically independent, and have the same variance
- Geometrically, the samples fall in equal-size hyperspherical clusters
- Decision boundary: hyperplane of d-1 dimension

$$\Sigma = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}, |\Sigma| = \sigma^{2d}, \Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix}$$

Linear Discriminant Function and Linear Machine

$\|\vec{x}-\vec{\mu}_i\|$: the Euclidean norm (distance)

$$\|\vec{x}-\vec{\mu}_i\|^2 = (\vec{x}-\vec{\mu}_i)^T(\vec{x}-\vec{\mu}_i)$$

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

$$= -\frac{\vec{x}^T\vec{x} - 2\vec{\mu}_i^T\vec{x} + \vec{\mu}_i^T\vec{\mu}_i}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(\vec{x}) = \frac{\vec{\mu}_i^T}{\sigma^2}\vec{x} - \frac{\vec{\mu}_i^T\vec{\mu}_i}{2\sigma^2} + \ln P(\omega_i)$$

Minimum-Distance Classifier

- When $P(\omega_i)$ are the same for all c classes, the discriminant function is actually measuring the minimum distance from each x to each of the c mean vectors

$$g_i(\vec{x}) = -\frac{\left\|\vec{x} - \vec{\mu}_i\right\|^2}{2\sigma^2}$$

Case 2: $\Sigma_i = \Sigma$

- The covariance matrices for all the classes are identical but not a scalar of identity matrix.
- Geometrically, the samples fall in hyperellipsoidal
- Decision boundary: hyperplane of d-1 dimension

$$g_i(\vec{x}) = \ln p(\vec{x} \mid \omega_i) + \ln P(\omega_i)$$

$$= -\frac{1}{2}\underbrace{\boxed{(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i})}_{\text{Squared Mahalanobis distance}} + \ln P(\omega_i)$$

Squared Mahalanobis distance

$$= \vec{\mu}_i^T \left(\Sigma^{-1}\right)^T \vec{x} - \frac{1}{2}\vec{\mu}_i^T \Sigma^{-1}\vec{\mu}_i + \ln P(\omega_i)$$

Case 3: Σ_i = arbitrary

- The covariance matrices are different from each category
- Quadratic classifier
- Decision boundary: hyperquadratic for 2-D Gaussian

$$g_i(\vec{x}) = \ln p(\vec{x} \mid \omega_i) + \ln P(\omega_i)$$

$$= -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= -\frac{1}{2}\vec{x}^T \Sigma_i^{-1} \vec{x} + \vec{\mu}_i^T \left(\Sigma_i^{-1}\right)^T \vec{x} - \frac{1}{2}\vec{\mu}_i^T \Sigma_i^{-1} \vec{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

AICIP RESEARCH

Questions

- What is a discriminant function?
- What is a multivariate Gaussian (or normal density function)?
- What is the covariance matrix and what is its dimension?
- What would the covariance matrix look like if the features are independent from each other?
- What would the covariance matrix look like if the features are independent from each other AND have the same spread in each dimension?
- What is minimum (Euclidean) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Euclidean) distance classifier?
- What is minimum (Mahalanobis) distance classifier? Is it a linear or quadratic classifier (machine)? What does the decision boundary look like?
- What are the assumptions made when using a minimum (Mahalanobis) distance classifier?
- What does the decision boundary look like for a quadratic classifier?
- What are the cost functions for the discriminant functions? And what is the optimization method used to find the best solution?

Multi-variate Gaussian

Linear and Quadratic Machines and their assumptions

Bayes Decision Rule

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

Maximum Posterior Probability

For a given x, if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then x belongs to class 1, otherwise, 2.

Discriminant Function

The classifier will assign a feature vector x to class ω_i if $g_i(x) > g_j(x)$

Case 1: Minimum Euclidean Distance (Linear Machine), $\Sigma_i=\sigma^2 I$

Case 2: Minimum Mahalanobis Distance (Linear Machine), $\Sigma_i = \Sigma$

Case 3: Quadratic classifier , Σ_i = arbitrary

All assuming Gaussian pdf