EE269
Signal Processing for Machine Learning

Lecture 8

Instructor : Mert Pilanci

Stanford University

February 4, 2019

Recap: Linear and Quadratic Discriminant Analysis

- Suppose $x[n] = [x_1, ...x_N] \sim N(\mu_k, \Sigma)$ when $y = k$
  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$
- K classes

$$f(x) = \arg \max_{k=1,\ldots,K} \pi_k g_k(x)$$

Scaled identity covariances $\Sigma_k = \sigma^2 I$

- Suppose $x[n] = [x_1, ... x_N] \sim N(\mu_k, \Sigma_k)$ when $y = k$
  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$
- K classes
- Decision boundary: hyperplane

$$w^T(x - x_0) = 0$$

$$\begin{aligned} w &= \mu_i - \mu_j \\ x_0 &= \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \log \frac{\pi_i}{\pi_j}(\mu_i - \mu_j) \end{aligned}$$

Scaled identity covariances $\Sigma_k = \sigma^2 I$

- Suppose $x[n] = [x_1, ...x_N] \sim N(\mu_k, \Sigma_k)$ when $y = k$
  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$
- K classes
- Decision boundary: hyperplane

$$w^T(x - x_0) = 0$$

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \log \frac{\pi_i}{\pi_j}(\mu_i - \mu_j)$$

- Hyperplane passes through the point x_0 and is orthogonal to w

Identical covariances $\Sigma_k = \Sigma$

- Suppose $x[n] = [x_1, ...x_N] \sim N(\mu_k, \Sigma)$ when $y = k$
  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$
- Decision boundary: hyperplane

$$w^T(x - x_0) = 0$$

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\log \frac{\pi_i}{\pi_j}}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Identical covariances $\Sigma_k = \Sigma$

- Suppose $x[n] = [x_1, ...x_N] \sim N(\mu_k, \Sigma)$ when $y = k$

  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$
- Decision boundary: hyperplane

$$w^T(x - x_0) = 0$$

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

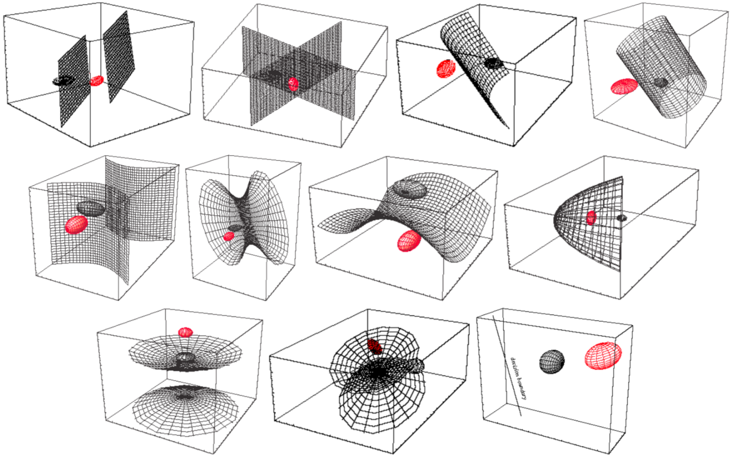
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\log \frac{\pi_i}{\pi_j}}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

- Hyperplane passes through x_0 but not necessarily orthogonal to the lines between the means

Quadratic Discriminant Analysis: Σ_k arbitrary

- Suppose $x[n] = [x_1, ...x_N] \sim N(\mu_k, \Sigma_k)$ when $y = k$
  $g_k(x) = P_{x|y=k} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$
- $h_k(x) = x^T W_k x + w_k^T x + w_{k0}$
- Classify as class k if $h_k(x) > h_{k'}(x) \quad \forall k' \neq k$
  $W_k = -\frac{1}{2}\Sigma_k^{-1}$
  $w_k = \Sigma_k^{-1}\mu_k$
  $w_{k0} = -\frac{1}{2}\mu_k^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$

Quadratic decision regions: hyperquadrics



Estimating parameters: univariate Gaussian

- Suppose $x_1, x_2, ...x_n$ i.i.d. $\sim N(\mu, \sigma^2)$
- Estimating means
  $\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_n$
- Estimating variances
  $\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_n - \mu_{ML})^2$

Estimating parameters: multivariate Gaussian

- Suppose $x_1, x_2, ...x_n$ i.i.d. $\sim N(\mu, \Sigma)$
- Estimating means

  $\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_n$

Estimating parameters: multivariate Gaussian

- Suppose $x_1, x_2, ...x_n$ i.i.d. $\sim N(\mu, \Sigma)$
- Estimating means

  $\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_n$
- Estimating covariances

  $\Sigma_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_n - \mu_{ML})(x_n - \mu_{ML})^T$

Linear vs Quadratic Discriminant Analysis

- LDA

  Estimate μ_k, for $k = 1..., K$ and Σ

  $Kn + \binom{n}{2} + n$ parameters

Linear vs Quadratic Discriminant Analysis

- LDA

  Estimate μ_k, for $k = 1..., K$ and Σ

  $Kn + \binom{n}{2} + n$ parameters

- QDA

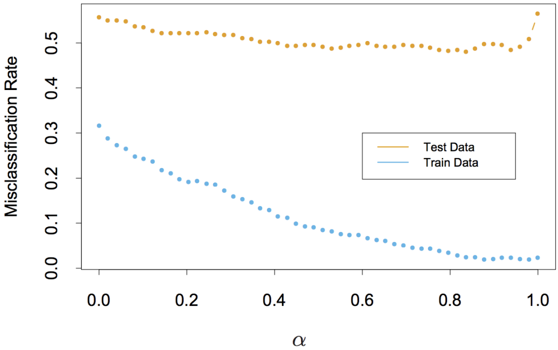
  Estimate μ_k, $\Sigma_k \quad$ for $k = 1..., K$

  $Kn + K\left(\binom{n}{2} + n\right)$ parameters

Regularized Linear Discriminant Analysis

- Maximum Likelihood Covariance estimate
  $\Sigma_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$
- Regularized estimate
  $\hat{\Sigma} = (1 - \alpha) \operatorname{diag}(\Sigma_{ML}) + \alpha \Sigma_{ML}$
- Diagonal Linear Discriminant Analysis ($\alpha = 0$)
  $\hat{\Sigma} = \operatorname{diag}(\Sigma_{ML})$

Regularized Discriminant Analysis on the Vowel Data



(slide credit: T. Hastie et al.)

Optimal basis change and dimension reduction

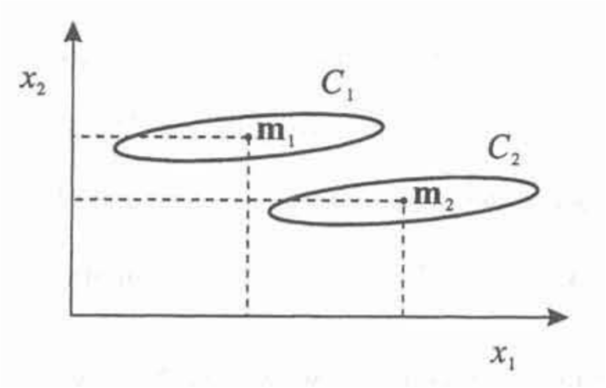
- Decision boundary $w^T(x - x_0) = 0$
  e.g., in LDA with equal covariances, $w = \Sigma^{-1}(\mu_i - \mu_j)$
- Classifies based on $w^T x \in \mathbb{R}$

Mean of the projected data

- $y = a^T x$

  $\mu_1 = \frac{1}{N_1} \sum_{i \in \text{ class } 1} x_i$

  $\mu_2 = \frac{1}{N_2} \sum_{i \in \text{ class } 2} x_i$



Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

  $\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}]$

  $\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}]$

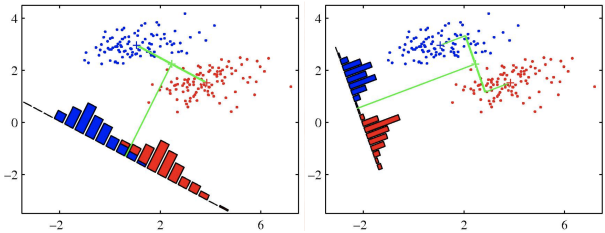
Fisher's LDA

- $\mu_k = \mathbb{E}[x \mid x \text{ comes from class k}]$
- $\Sigma_k = \mathbb{E}(x - \mu_k)(x - \mu_k)^T \mid x \text{ comes from class k}]$
- classify using a scalar feature $y = a^T x$

  $\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}]$

  $\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}]$

$$\max_a \frac{(\beta_1 - \beta_2)^2}{\sigma_1^2 + \sigma_2^2}$$



Fisher's LDA

$$\beta_k = \mathbb{E}[y \mid x \text{ comes from class k}] = a^T \mu_k$$
$$\sigma_k^2 = \mathbb{E}[(y - \beta_k)^2 \mid x \text{ comes from class k}] =$$
$$\mathbb{E}[(a^T(x - \mu_k))^2] = \mathbb{E}[(a^T(x - \mu_k)(x - \mu_k)^T a] = a^T \Sigma_k a$$

$$\begin{aligned} \max_a \frac{(\beta_1 - \beta_2)^2}{\sigma_1^2 + \sigma_2^2} &= \max_a \frac{(a^T(\mu_1 - \mu_2))^2}{a^T(\Sigma_1 + \Sigma_2)a} \\ &= \max_a \frac{a^T Q a}{a^T P a} \end{aligned}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

Maximizing quadratic forms

$$\max_a \frac{a^T Q a}{a^T a}$$

Maximizing quadratic forms

$$\max_a \frac{a^T Q a}{a^T a}$$

- Eigenvalue Decomposition $Q = U \Lambda U^T$
- Change of basis $b = U^T a$, i.e., $Ub = a$

$$\begin{aligned}\max_a \frac{a^T U \Lambda U^T a}{a^T a} &= \max_b \frac{b^T \Lambda b}{b^T U^T U b} \\ &= \max_b \frac{b^T \Lambda b}{b^T b}\end{aligned}$$

Maximizing quadratic forms

$$\max_a \frac{a^T Q a}{a^T a}$$

- Eigenvalue Decomposition $Q = U \Lambda U^T$
- Change of basis $b = U^T a$, i.e., $Ub = a$

$$\begin{aligned}\max_a \frac{a^T U \Lambda U^T a}{a^T a} &= \max_b \frac{b^T \Lambda b}{b^T U^T U b} \\ &= \max_b \frac{b^T \Lambda b}{b^T b}\end{aligned}$$

- Optimum is given by $b = \delta[n - k^*]$ where

$$k^* = \arg\max_k \Lambda_{kk} = 1$$

Solution: $a = u_1$ maximal eigenvector, i.e., $Q u_1 = \lambda_1 u_1$
Optimal value : λ_1

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Theorem (Simultaneous Diagonalization)

  Let $P, Q \in \mathbb{R}^{n \times n}$ real symmetric matrices, and P is positive definite, then there exists a matrix V such that

$$V^T P V = I$$
$$V^T Q V = \Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$$

  where V, Λ satisfies the generalized eigenvalue equation:

$$Q v_i = \lambda_i P v_i$$

Maximizing quadratic forms: two quadratics

- Theorem (Simultaneous Diagonalization)

  Let $P, Q \in \mathbb{R}^{n \times n}$ real symmetric matrices, and P is positive definite, then there exists a matrix V such that

$$V^T P V = I$$
$$V^T Q V = \Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$$

  where V, Λ satisfies thegeneralized eigenvalue equation:

$$Q v_i = \lambda_i P v_i$$

  Proof: Let $P = U_P \Lambda_P U_P^T$ be its Eigenvalue Decomposition

  $V' = U_P \Lambda_P^{-\frac{1}{2}}$ will only diagonalize P

  Let ${V'}^T Q V' = U' \Lambda' {U'}^T$ be its EVD

  Set $V = V' U'$ $\square$

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Let V and Λ satisfy the generalized eigenvalue equation

  $$Qv_i = \lambda_i P v_i$$

  Basis change $a = Vb$, i.e., $b = V^T a$

Maximizing quadratic forms: two quadratics

$$\max_a \frac{a^T Q a}{a^T P a}$$

- Let V and Λ satisfy the generalized eigenvalue equation

  $$Qv_i = \lambda_i P v_i$$

  Basis change $a = Vb$, i.e., $b = V^T a$

  $$\max_b \frac{b^T V^T Q V b}{b^T V^T P V b} = \max_b \frac{b^T \Lambda b}{b^T b}$$

- Solution: $a = v_1$, maximal generalized eigenvector

  Optimal value: λ_1 maximum generalized eigenvalue

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

$$P^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T a = \lambda a$$

Fisher's LDA

$$\max_a \frac{a^T Q a}{a^T P a}$$

where $Q = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $P = \Sigma_1 + \Sigma_2$.

- Solution: $Qa = \lambda P a$, therefore $P^{-1} Q a = \lambda a$

$$P^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T a = \lambda a$$

$a = \text{constant} \times P^{-1}(\mu_1 - \mu_2)$

can be normalized as $a := \frac{P^{-1}(\mu_1 - \mu_2)}{||P^{-1}(\mu_1 - \mu_2)||_2}$