Fundamentals of Machine Learning

LINEAR MODELS

Amit K Roy-Chowdhury

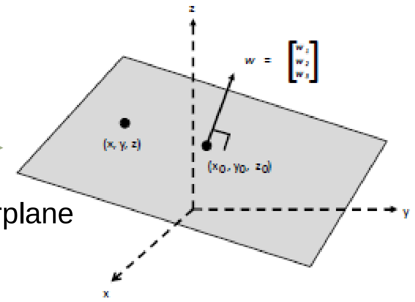
Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee

Linear Classifier

The prediction can be written as

$$f(x) = \mathbb{I}\left(p(y=1|x) > p(y=0|x)\right) = \mathbb{I}\left(\log \frac{p(y=1|x)}{p(y=0|x)} > 0\right) = \mathbb{I}\left(a > 0\right)$$

$$f(x;\theta) = b + \boldsymbol{w}^T \boldsymbol{x}$$



hyperplane

normal vector $w \in \mathbb{R}^D$ and an offset $b \in \mathbb{R}$ from the origin

This linear hyperplane separate 3d space into half □ decision boundary

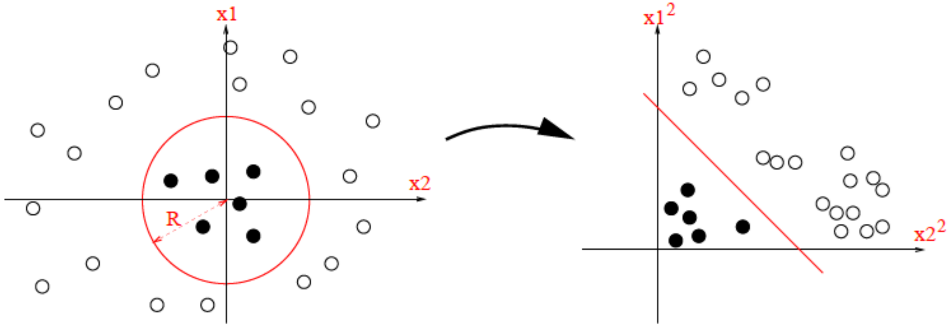
Non Linear Classifier

Transform input features in suitable way

$$\phi(x_1, x_2) = [1, x_1^2, x_2^2]$$

$$\boldsymbol{w} = [-R^2, 1, 1]. \text{ Then } \boldsymbol{w}^\top \phi(\boldsymbol{x}) = x_1^2 + x_2^2 - R^2$$

Decision boundary (where f(x) = 0) defines a circle with radius R



Outline

- Logistic Regression
- Linear Regression
- **Linear Discriminant Analysis**
- Naïve Bayes

Linear Discriminant Analysis

class conditional

prior

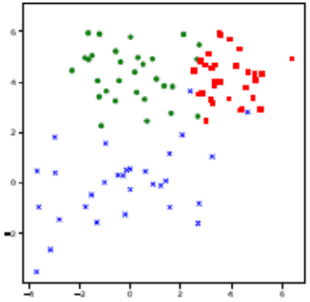
$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y=c;\boldsymbol{\theta})p(y=c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y=c';\boldsymbol{\theta})p(y=c';\boldsymbol{\theta})}$$

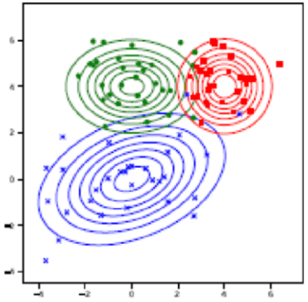
posterior

Linear Discriminant Analysis: $\log p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + \text{const}$

Gaussian Discriminant Analysis: $p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$

$$\Rightarrow p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c) \quad \text{where } \pi_c = p(y=c)$$





Linear Discriminant Analysis

class conditional

prior

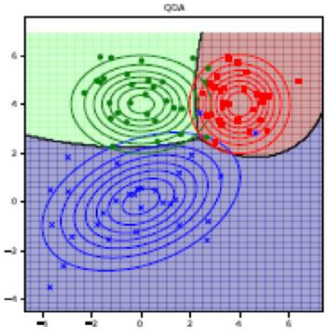
$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y=c;\boldsymbol{\theta})p(y=c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y=c';\boldsymbol{\theta})p(y=c';\boldsymbol{\theta})}$$

posterior

Gaussian Discriminant Analysis: $p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$

$$\Rightarrow p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \quad \text{where } \pi_c = p(y=c)$$

Discriminant Function: $\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \log \pi_c - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const}$



Linear Discriminant Analysis

class conditional

prior

$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}|y=c;\boldsymbol{\theta})p(y=c;\boldsymbol{\theta})}{\sum_{c'} p(\boldsymbol{x}|y=c';\boldsymbol{\theta})p(y=c';\boldsymbol{\theta})}$$

posterior

Gaussian Discriminant Analysis: $p(\boldsymbol{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$

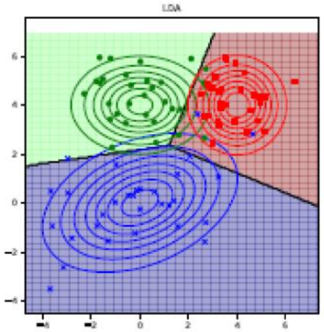
$$\Rightarrow p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \propto \pi_c \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \quad \text{where } \pi_c = p(y=c)$$

Discriminant Function: $\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = \log \pi_c - \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const}$

$\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$

$$\begin{aligned}\log p(y=c|\boldsymbol{x},\boldsymbol{\theta}) &= \log \pi_c - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \text{const} \\ &= \underbrace{\log \pi_c - \frac{1}{2}\boldsymbol{\mu}_c^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c}_{\gamma_c} + \boldsymbol{x}^{\mathsf{T}}\underbrace{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c}_{\boldsymbol{\beta}_c} + \underbrace{\text{const} - \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}_{\kappa}\end{aligned}$$

LInear Discriminant Analysis: $= \gamma_c + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}_c + \kappa$



Interpretation of LDA

Uniform prior over classes

$$\hat{y}(\boldsymbol{x}) = \operatorname*{argmax}_c \log p(y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \operatorname*{argmin}_c (\boldsymbol{x} - \boldsymbol{\mu}_c)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)$$

nearest centroid classifier or nearest class mean classifier

Fisher's LDA

Reduce feature dimensionality (PCA!) and classify.

find the matrix $\mathbf{W}$ such that the low-dimensional data can be classified as well as possible

$x \in \mathbb{R}^D$

$z \in \mathbb{R}^K$

$$z_n = \mathbf{W}x_n$$

data points

$$m_c = \frac{1}{N_c}\sum_{n:y_n=c} z_n$$

mean for class c

$$m = \frac{1}{N}\sum_{c=1}^{C} N_c m_c$$

overall mean

$$\tilde{\mathrm{S}}_W = \sum_{c=1}^{C}\sum_{n:y_n=c}(z_n - m_c)(z_n - m_c)^\mathsf{T}$$

$$\tilde{\mathrm{S}}_B = \sum_{c=1}^{C} N_c(m_c - m)(m_c - m)^\mathsf{T}$$

scatter matrices

maximize objective function

$$J(\mathbf{W}) = \frac{|\tilde{\mathrm{S}}_B|}{|\tilde{\mathrm{S}}_W|} = \frac{|\mathbf{W}^\mathsf{T}\mathrm{S}_B\mathbf{W}|}{|\mathbf{W}^\mathsf{T}\mathrm{S}_W\mathbf{W}|}$$

Leads to a generalized eigenvalue problem – advanced reading

FLDA – 2 classes

$$S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^\mathsf{T}$$

$$J(w) = \frac{w^\mathsf{T} S_B w}{w^\mathsf{T} S_W w}$$

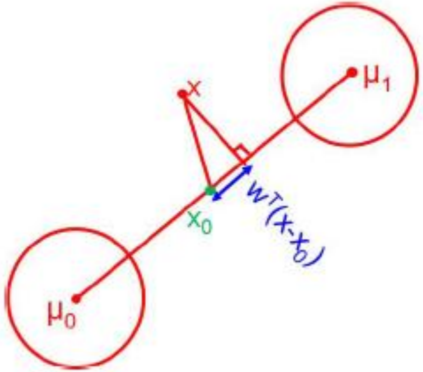
$$S_W = \sum_{n:y_n=1} (x_n - \mu_1)(x_n - \mu_1)^\mathsf{T} + \sum_{n:y_n=2} (x_n - \mu_2)(x_n - \mu_2)^\mathsf{T}$$

$$\mu_1 = \frac{1}{N_1} \sum_{n:y_n=1} x_n, \ \mu_2 = \frac{1}{N_2} \sum_{n:y_n=2} x_n$$

Take derivative wrt w

$$S_B w = \lambda S_W w \quad \text{where} \quad \lambda = \frac{w^\mathsf{T} S_B w}{w^\mathsf{T} S_W w}$$

Generalized eigenvalue problem, becomes regular eigenvalue problem if $S_W^{-1} S_B w = \lambda w$



Interpretation:

Let $m_k = w^\mathsf{T} \mu_k$ be the projection of each mean onto the line w.

$$S_B w = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^\mathsf{T} w = (\mu_2 - \mu_1)(m_2 - m_1)$$

$$\lambda \, w = S_W^{-1} (\mu_2 - \mu_1)(m_2 - m_1)$$

$$w \propto S_W^{-1} (\mu_2 - \mu_1)$$ w is proportional to the vector that joins the class means

Outline

- Logistic Regression
- Linear Regression
- Linear Discriminant Analysis
- Naïve Bayes

Naïve Bayes

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

Naive Bayes Assumption:

$$P(\mathbf{x}|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y), \text{ where } x_\alpha = [\mathbf{x}]_\alpha \text{ is the value for feature } \alpha$$

$$\begin{aligned}
h(\mathbf{x}) &= \operatorname*{argmax}_y P(y|\mathbf{x}) \\
&= \operatorname*{argmax}_y \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \\
&= \operatorname*{argmax}_y P(\mathbf{x}|y)P(y) && (P(\mathbf{x}) \text{ does not depend on } y) \\
&= \operatorname*{argmax}_y \prod_{\alpha=1}^{d} P(x_\alpha|y)P(y) && (\text{by the naive Bayes assumption}) \\
&= \operatorname*{argmax}_y \sum_{\alpha=1}^{d} \log(P(x_\alpha|y)) + \log(P(y)) && (\text{as log is a monotonic function})
\end{aligned}$$

Simple Example

- Given N1 emails which are spam and N2 not spam; p(S) = N1/(N1+ N2); p(NS) = N2/N1+N2
- Consider some words that occur in each category with some frequency: p(w1|S), p(w2|S),P(w3|S),……;
  p(w1|NS), p(w2|NS),…..
- Say you observe W = {w1, w5, w7} in an email. Is it spam or not spam?

- p(S|W) $\propto$ p(S) x p(W|S) – now use conditional independence
- P(NS|W) $\propto$ p(NS) x p(W|NS) – now use conditional independence

- What happens if one word, say w7, never occurred in the training data for NS?

- What is it we needed to know to calculate this?
  - Number of times each word occurred in each class as a fraction of all the words in that class.