Lecture 9: Classification (Textbook 4.4 and 4.5)

Overview

Recall that Bayes theorem provides

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

For each class $k \in [K]$,

- π_k is easily estimated using the proportion of observation in classe k.
- f_k is hard to estimate (p dimensional density function)

- LDA : f_k is the density of $\mathcal{N}_p(\mu_k, \Sigma)$
- QDA : f_k is the density of $\mathcal{N}_p(\mu_k, \Sigma_k)$
- Naive Bayes : the p predictors are independent ($f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$)

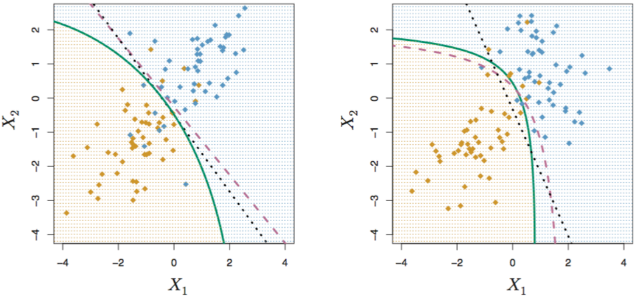
Lot of technical details given during the lecture !

Quadratic Discriminant Analysis

In QDA, the Bayes classifier assigns an observation $X = x$ to the class for which

$$\delta_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log\pi_k - \frac{1}{2}\log|\Sigma_k|.$$

is largest. So, the decision boundary is nonlinear (quadratic).



The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries under two scenarios.

Naive Bayes

Assumes features are independent in each class.

Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Under Gaussian distributions, naive Bayes assumes each Σ_k is diagonal. The decision boundary is determined by

$$\delta_k(x) = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k.$$

- It is easy to extend it to mixed features (quantitative and categorical).

- Despite strong assumptions, naive Bayes often produces good classification results.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c1x_1 + ... + c_p x_p,$$

which has the same form as logistic regression.

The difference is in how the parameters are estimated.

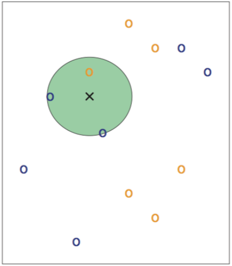
- Logistic regression uses the conditional likelihood based on $P(Y|X)$ (known as discriminative learning).
- LDA uses the full likelihood based on $P(X, Y)$ (known as generative learning).
- Despite these differences, in practice the results are often very similar.

K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) classifier directly estimates $P(Y = j|X = x_0)$ by

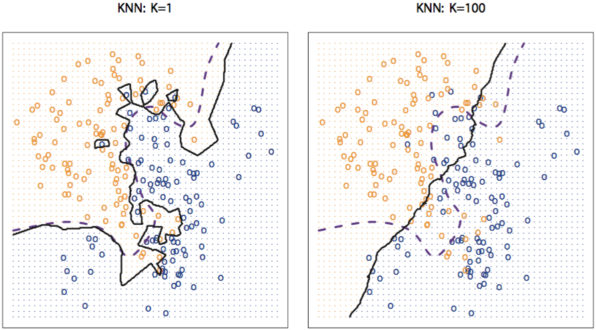
$$\frac{1}{K}\sum_{i \in N_0} I(y_i = j),$$

where N_0 is the set of K points in the training data that are closest to x_0. KNN estimate $P(Y = j|X = x_0)$ as the fraction of points with label j in N_0.



(KNN with $K = 3$).

Effect of K



With K $=$ 1, the decision boundary is overly flexible, while with K $=$ 100 it is not sufficiently flexible. Again, this represents the bias variance trade-off. The Bayes decision boundary is shown as a purple dashed line.

Comparison

K classes
An obseravtion x is assigned to the class that maximizes $\mathbb{P}[Y = k \ X = x]$.
It is similar than asssuming class K is the baseline and maximizing the log odds

$$\log\left[\frac{\mathbb{P}[Y = k \ X = x]}{\mathbb{P}[Y = K \ X = x]}\right]$$

- LDA: log odds is LINEAR in x
- QDA: log odds is QUADRATIC in x
- Naive Bayes: log odds is a generalized additive model

Comparison bis

- LDA is a special case of QDA
- LDA is a special case of Naive Bayes (not trivial !)
- QDA is NOT a special case of Naive Bayes (and vice versa)

Which is better ?

- LDA outperforms MLR (Multinomial logistic regression) when Gaussian assumption holds
- KNN dominates LDA and MLR when the decision boundary is non linear and $n \gg p$
- QDA dominates LDA and MLR when the decision boundary is non linear and $n \gtrsim p$
- KNN doesn't tell which regressor is important

Read the textbok if you did not attend the lectures !