MONASH University

ETC3250

Business Analytics

Week 5
Comparison of classifiers

21 August 2017

Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3,7	Tas & David
4	Classification	4	Souhaib
5	Classification	4, 9	Souhaib
	Comparison of classifiers		Souhaib
	Support vector machines		Souhaib
6	Resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced learning methods	8	Souhaib
	Semester break		
10	Clustering	10	Souhaib
11	Visualization		Souhaib
12	Data wrangling		Souhaib

Optimal classifier

The optimal classifier (also called Bayes classifier) at $\boldsymbol{x}$ in terms of expected error rate, i.e the classifier which minimizes $\mathbb{E}[I(Y \neq C(\boldsymbol{x}))]$ is given by

$$C(\boldsymbol{x}) = j \quad \text{if } p_j(\boldsymbol{x}) = \max\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x}), \dots, p_K(\boldsymbol{x})\}$$

where

$$p_k(\boldsymbol{x}) = \Pr(Y = k \mid \boldsymbol{X} = \boldsymbol{x}), \qquad k = 1, 2, \dots, K.$$

$\rightarrow$ In practice, we do not know $p_k(\boldsymbol{x})$; we only observe $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ where $(y_i, \boldsymbol{x}_i) \sim \Pr(Y, \boldsymbol{X})$.

Classification methods

- K-nearest neighbors (KNN)
- Logistic regression
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)

KNN Classifier

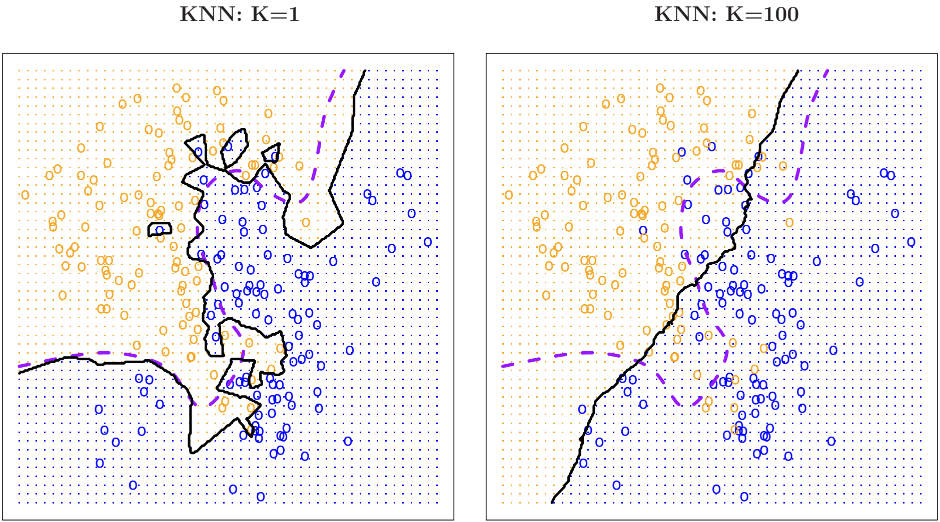
One of the simplest classifiers. Given a test observation x_0:

- Find the K nearest points to x_0 in the training data: $\mathcal{N}_0$.
- Estimate conditional probabilities

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

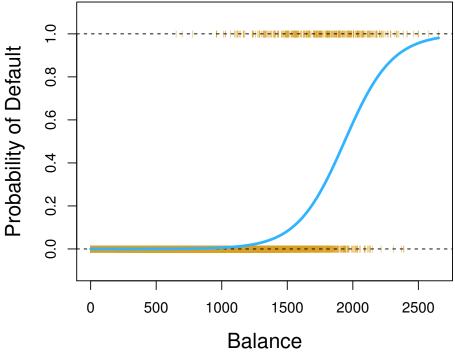
- Classify x_0 to class with largest probability.

KNN Classifier



Logistic regression

$$p(X) = P(Y = 1|X) = \text{logistic}(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$
$$\rightarrow log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$$



Linear/Quadratic Discriminant Analysis

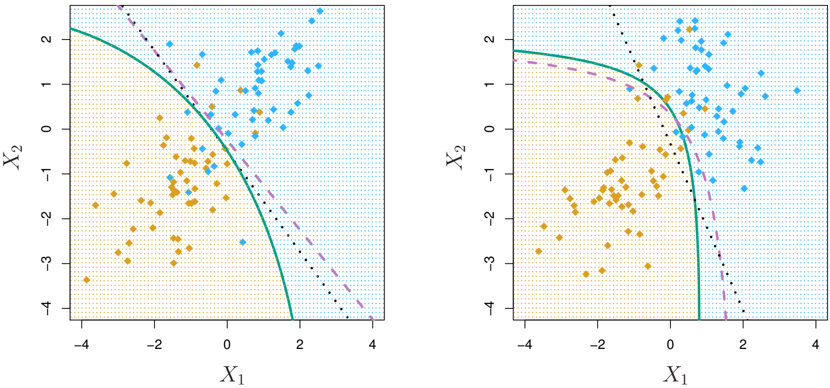
Using Bayes' therorem:

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

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

- Linear Discriminant Analysis (LDA)
  - Observations from the kth class: $X \sim N(\mu_k, \sigma^2)$
- Quadratic Discriminant Analysis (QDA)
  - Observations from the kth class: $X \sim N(\mu_k, \sigma_k^2)$

Linear/Quadratic Discriminant Analysis



- Bayes (purple dashed)
- QDA (green solid)
- LDA (black dotted)

Logistic regression and LDA

- Logistic regression
  - $log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x$
  - β_0 and β_1 estimated using maximum likelihood
- Linear Discriminant Analysis
  - $log(\frac{p_1(x)}{1-p_1(x)}) = c_0 + c_1 x$
  - c_0 and c_1 computed using the estimated mean and variance of a normal distribution

→ Both logistic regression and LDA produce **linear decision boundaries**.

→ However, they make **different assumptions** and use a different fitting procedure

KNN Classifier

- Nonparametric approach: **no assumptions** about the shape of the decision boundary
- We can expect KNN to dominate LDA and logistic regression when the decision boundary is **highly non-linear**
- KNN does not tell us which predictors are **important**; No table of coefficients as in logistic regression

QDA Classifier

- QDA serves as a **compromise** between the non-parametric KNN method and the linear LDA and logistic regression approaches
- Since QDA assumes a **quadratic decision boundary**, it can accurately model a wider range of problems than can the linear methods.
- QDA is **less flexible than KNN** but can perform better in the presence of a **limited number of training observations** because it does make some assumptions about the form of the decision boundary

Which classification method?

- Binary or multi-class classification?
- How many training examples do we have?
- What is the dimensionality of the problem?
- How many categorical variables do we have?
- Are features independent?
- Do we expect the classes to be linearly separable?
- Any requirements in terms of computational time/performance/memory usage?
- Importance of interpretability?

Empirical comparison of classifiers

- We compare the following classifiers: KNN-1, KNN-CV, LDA, Logistic and QDA
- We consider **six different scenarios** for the data generating process
- Scenarios 1-3 are **linear**, and scenarios 4-6 are **nonlinear**
- In each scenario, we generate 100 **random training data sets**. For each of these training sets, we fit each model to the data and compute the test error rate on a **large test set**

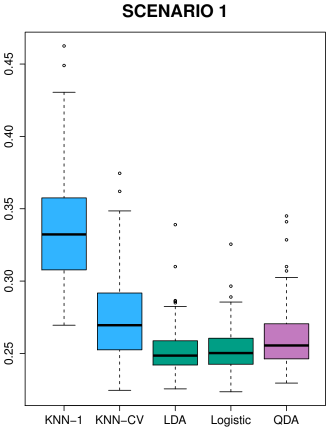
Scenario 1

There were 20 training observations in each of two classes. The observations within each class were uncorrelated random normal variables with a different mean in each class.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 1

There were 20 training observations in each of two classes. The observations within each class were uncorrelated random normal variables with a different mean in each class.



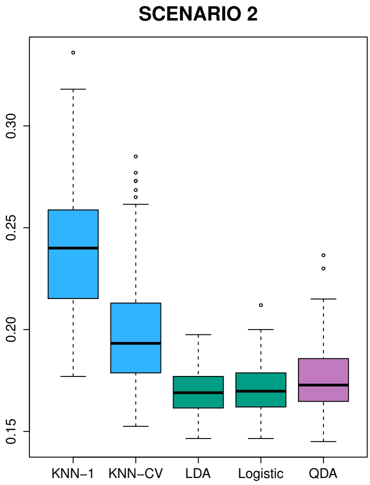
Scenario 2

Details are as in Scenario 1, except that within each class, the two predictors had a correlation of -0.5.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 2

Details are as in Scenario 1, except that within each class, the two predictors had a correlation of -0.5.



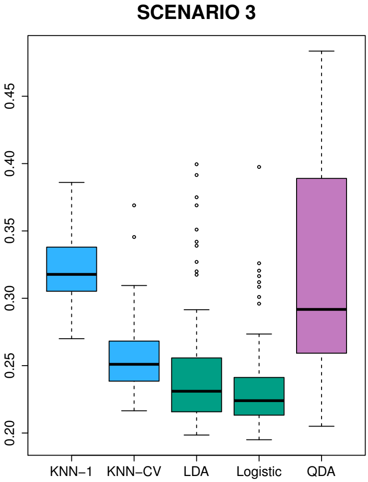
Scenario 3

We generated X_1 and X_2 from the t-distribution, with 50 observations per class.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 3

We generated X_1 and X_2 from the t-distribution, with 50 observations per class.



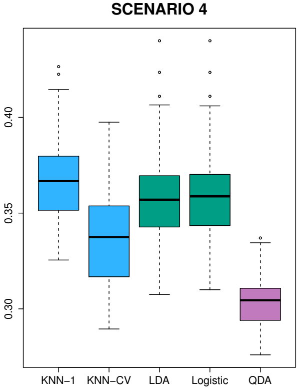
Scenario 4

The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of -0.5 between the predictors in the second class.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 4

The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of -0.5 between the predictors in the second class.



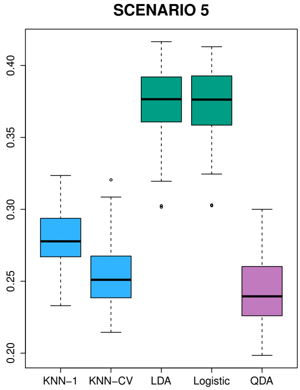
Scenario 5

Within each class, the observations were generated from a normal distribution with uncorrelated predictors. However, the responses were sampled from the logistic function using X_1^2, X_2^2 and $X_1 \times X_2$ as predictors.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 5

Within each class, the observations were generated from a normal distribution with uncorrelated predictors. However, the responses were sampled from the logistic function using X_1^2, X_2^2 and $X_1 \times X_2$ as predictors.



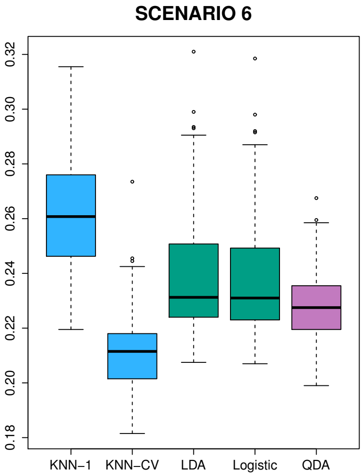
Scenario 6

Details are as in the previous scenario, but the responses were sampled from a more complicated non-linear function.

KNN-1, KNN-CV, LDA, Logistic and QDA?

Scenario 6

Details are as in the previous scenario, but the responses were sampled from a more complicated non-linear function.



Summary

- When the true decision boundaries are linear, LDA and logistic regression will perform well
- When the boundaries are moderately non-linear, QDA may give better results
- For more complicated boundaries, a non-parametric approach such as KNN can be superior
- Do not forget the importance of other criteria: number of samples and predictors, computational time, interpretability, etc.
- In many data analytics competitions, tree-based methods such as Boosting and Random Forests are often among the best methods (later).