MONASH University

ETC3250

Business Analytics

Week 5
Comparison of classifiers

24 September 2015

Outline

Week	Topic	Chapter	Lecturers
1	Introduction to business analytics & R	1	Rob,Souhaib
2	Statistical learning	2	Rob,Souhaib
3	Regression for prediction	3	Rob
4	Resampling	5	Rob, Souhaib
5	Dimension reduction	6,10	Rob, Souhaib
6	Visualization		Di
7	Visualization		Di
8	Classification	4	Souhaib, Di
9	Classification	4,9	Di, Souhaib
-	Semester Break		
10	Advanced classification	8	Di
11	Advanced regression	6	Di
12	Clustering	10	Di

Optimal classifier

The Bayes classifier is the **optimal classifier** under the error rate:

$$E[I(Y \neq \hat{f}(X))]$$

The Bayes classifier at x is given by

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

where

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

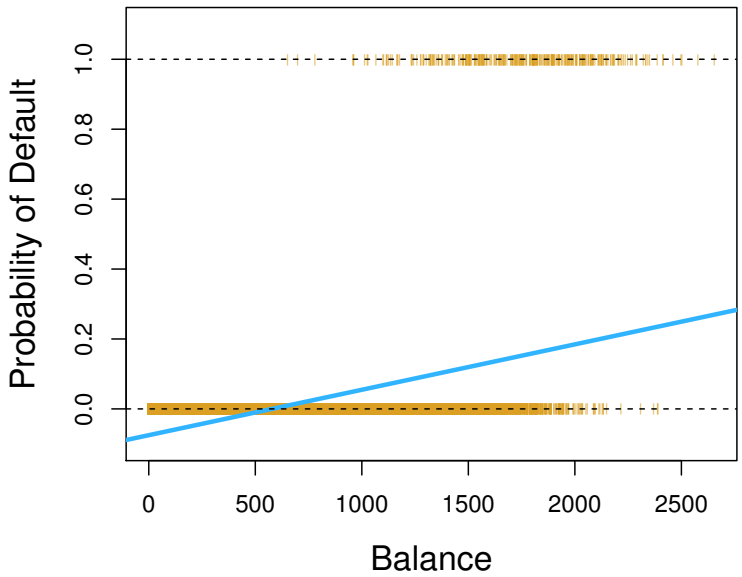
Logistic regression

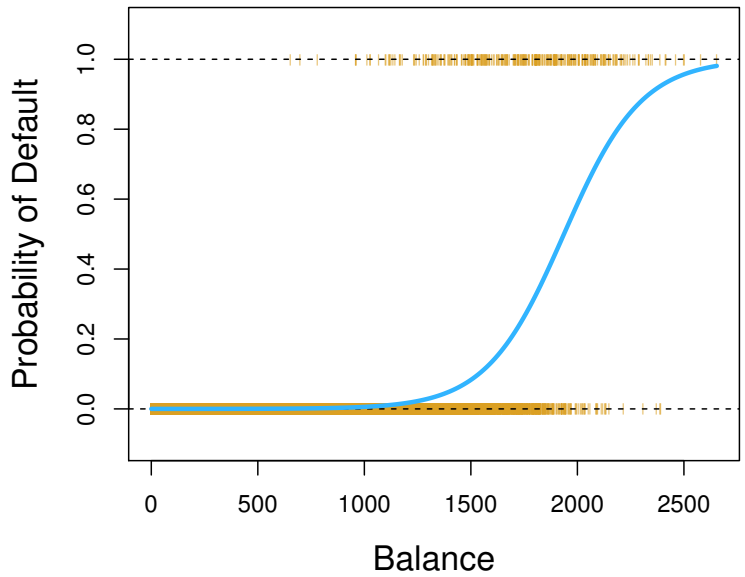
$$p(X) = P(Y = 1|X)$$

Linear reg. $p(X) = \beta_0 + \beta_1 X$

Logistic reg. $p(X) = \text{logistic}(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$

$$\rightarrow log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$





Linear/Quadratic Discriminant Analysis

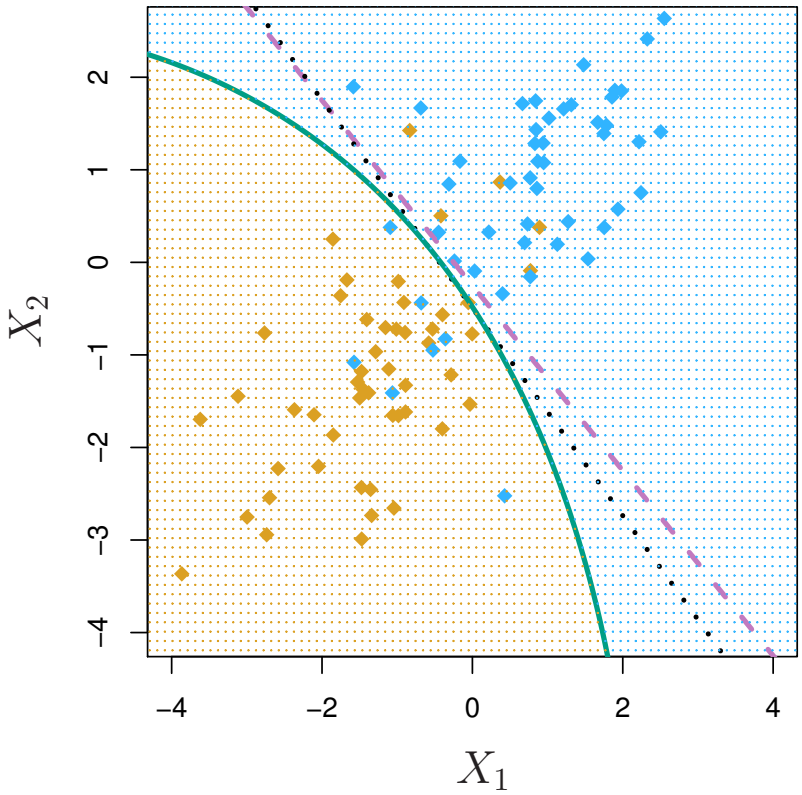
- Linear Discriminant Analysis (LDA)
  - Observations from the kth class: $X \sim N(\mu_k, \Sigma)$

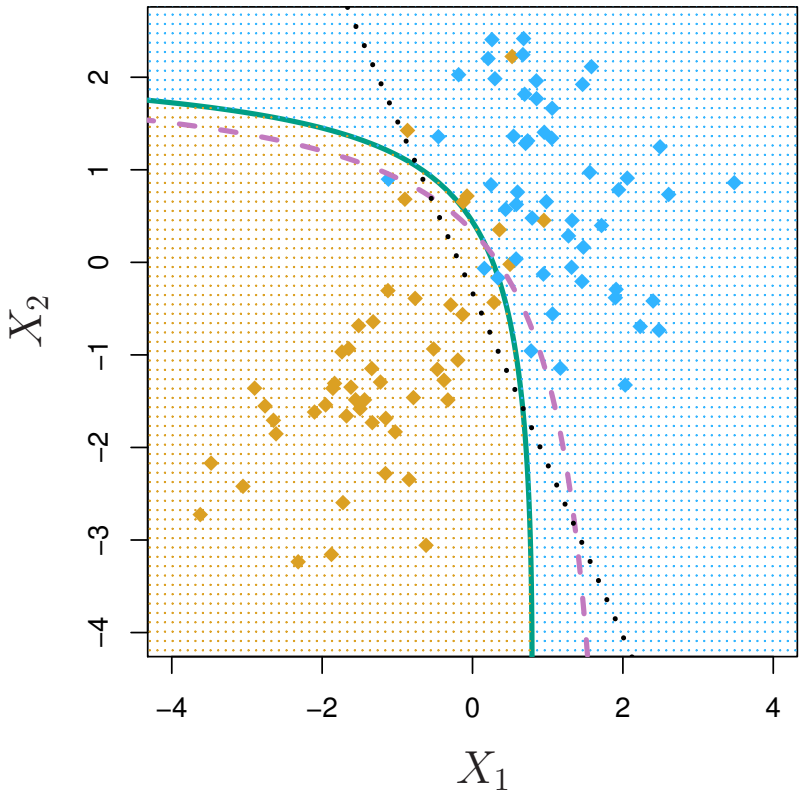
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

- Quadratic Discriminant Analysis (QDA)
  - Observations from the kth class: $X \sim N(\mu_k, \boldsymbol{\Sigma_k})$

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

Linear/Quadratic Discriminant Analysis





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. **Anything else?**

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

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

$\rightarrow$ Both logistic regression and LDA produce linear decision boundaries. **Anything else?**

$\rightarrow$ However, they make different assumptions and use a different fitting procedure

kNN Classifier

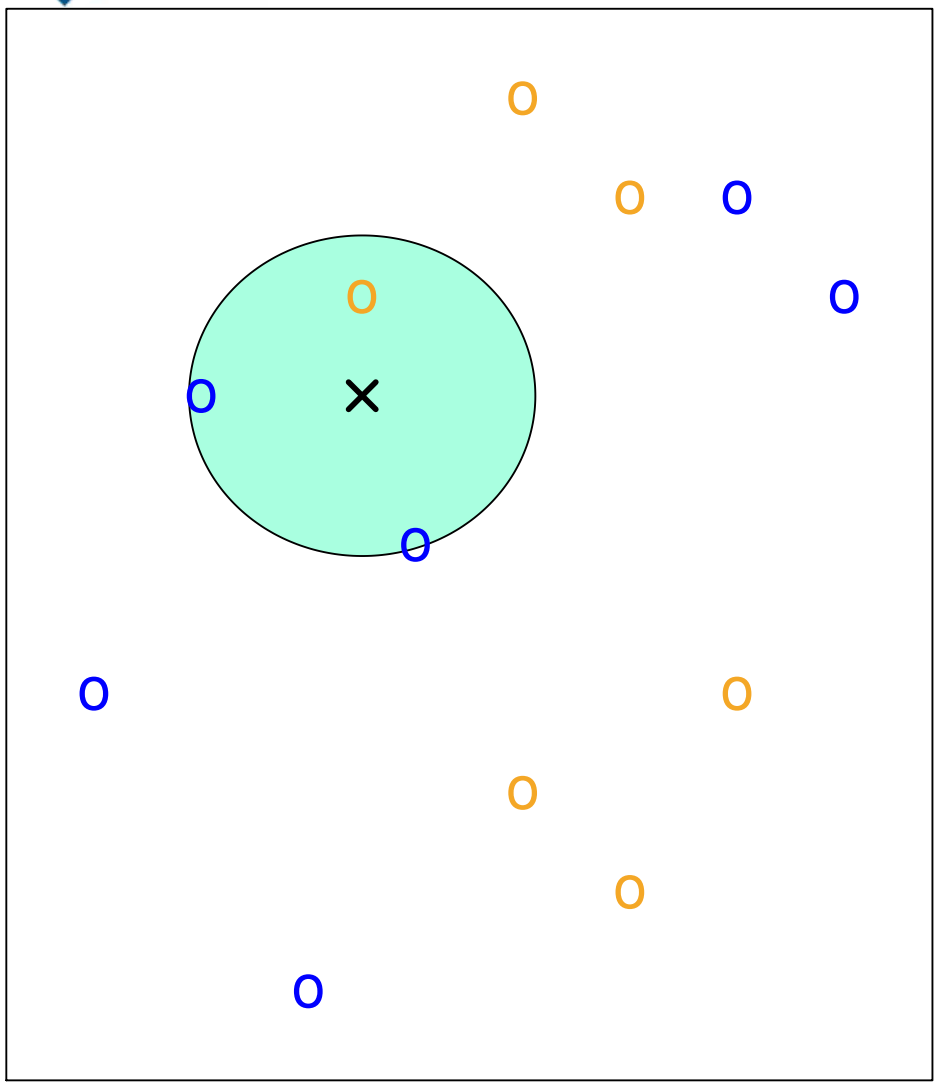
One of the simplest classifiers. Given a test observation x_0:

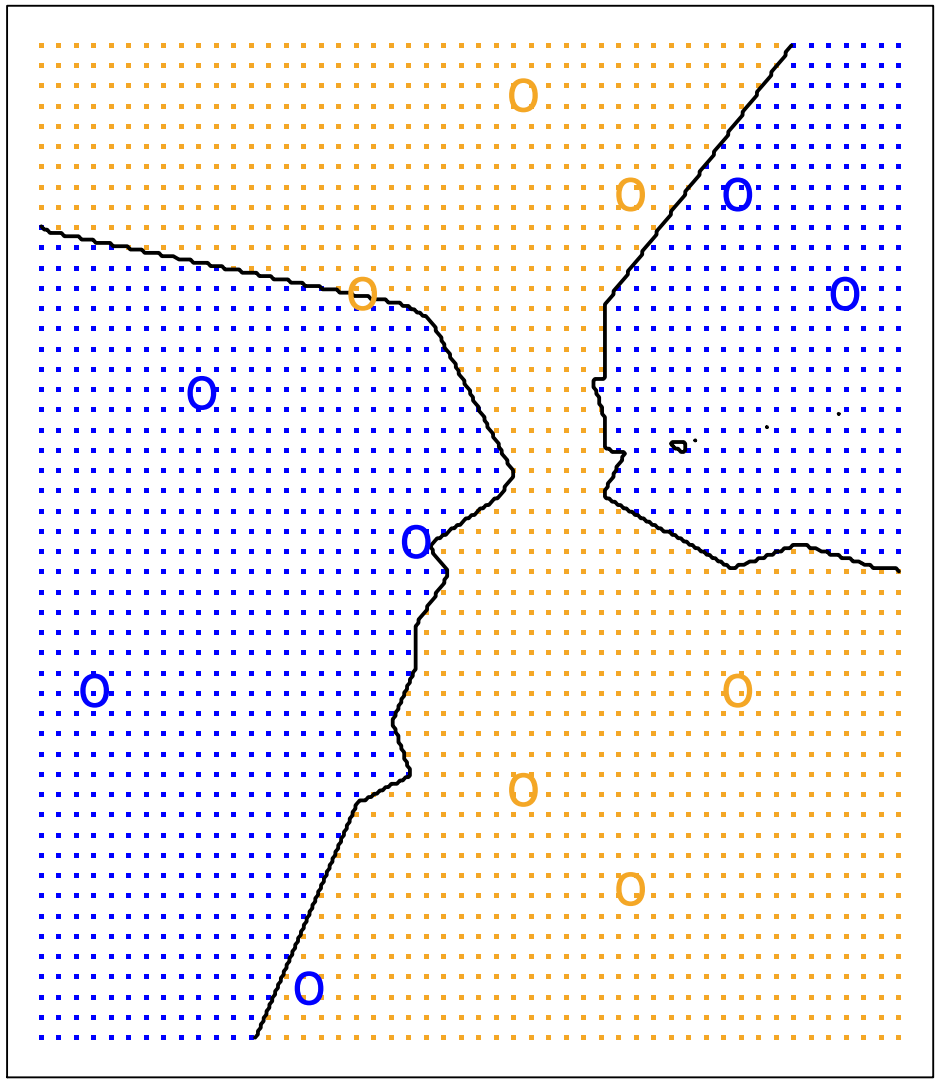
- Find the K nearest points to x_0 in the training data: $\mathcal{N}_0$.
- Estimate conditional probabilities $\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$.
- Classify x_0 to class with largest probability.

$\rightarrow$ Nonparametric approach: no assumptions about the shape of the decision boundary

$\rightarrow$ No table of coefficients as in logistic regression

kNN Classifier

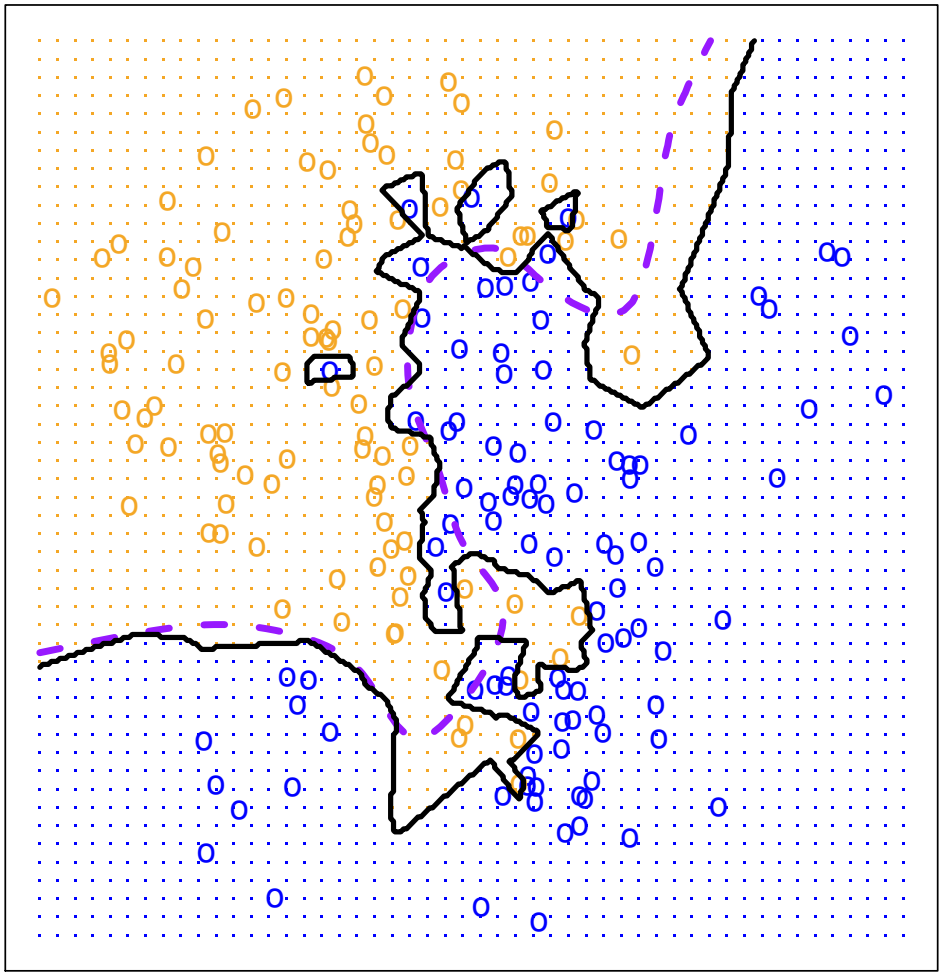




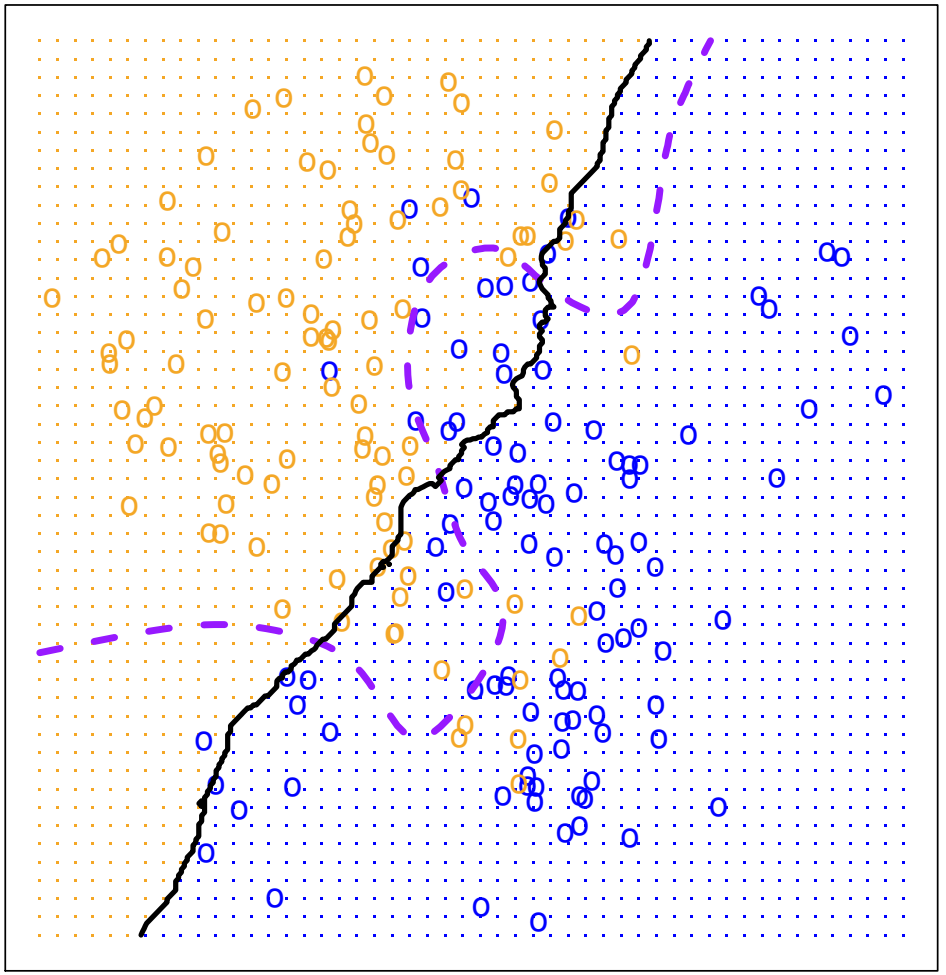
$K = 3$.

kNN Classifier

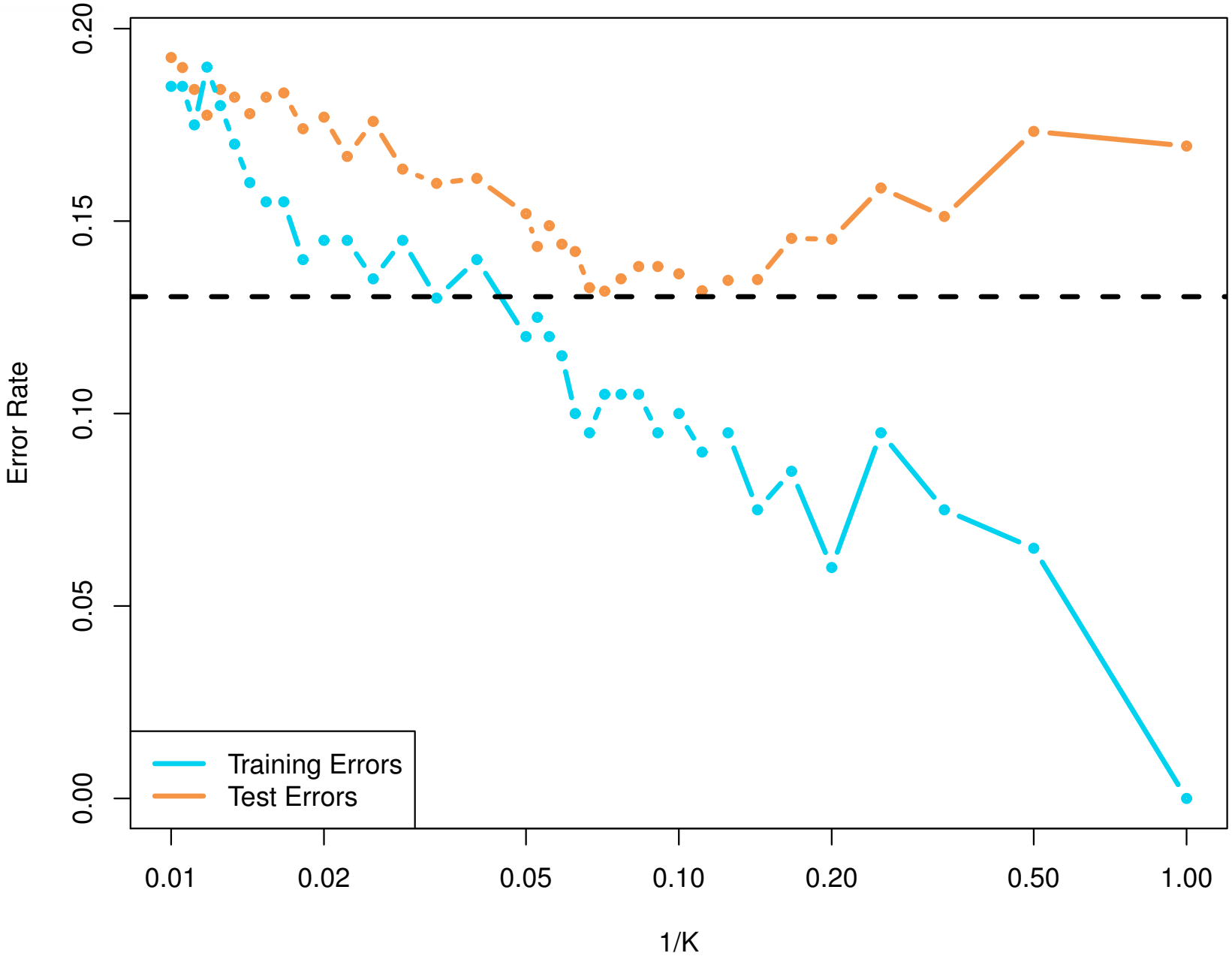
KNN: K=1



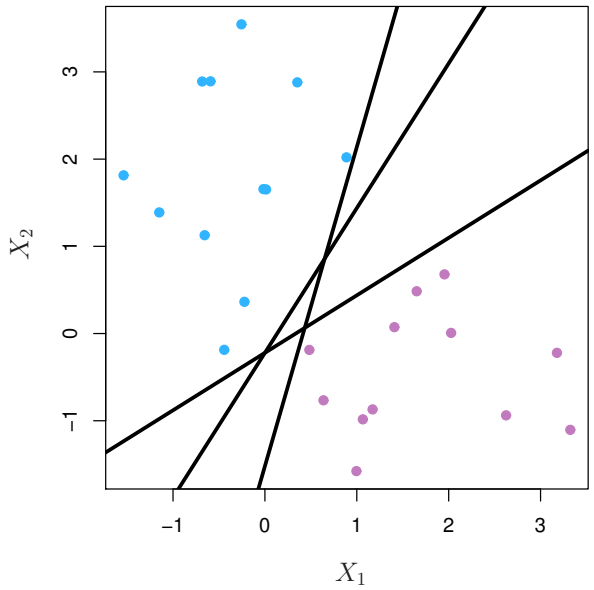
KNN: K=100

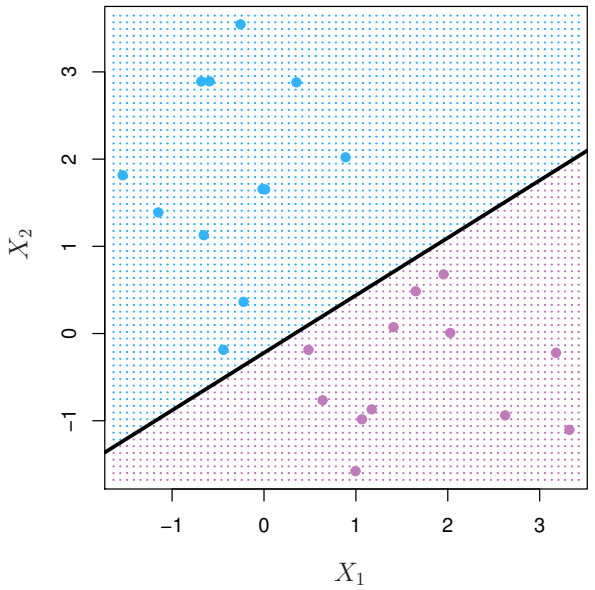


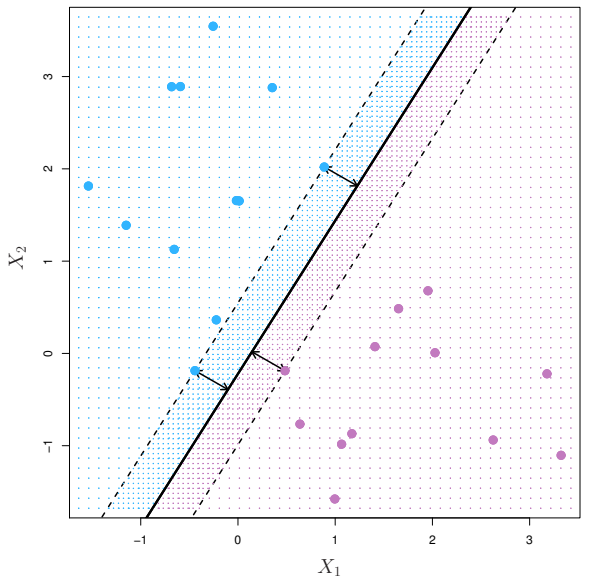
kNN Classifier

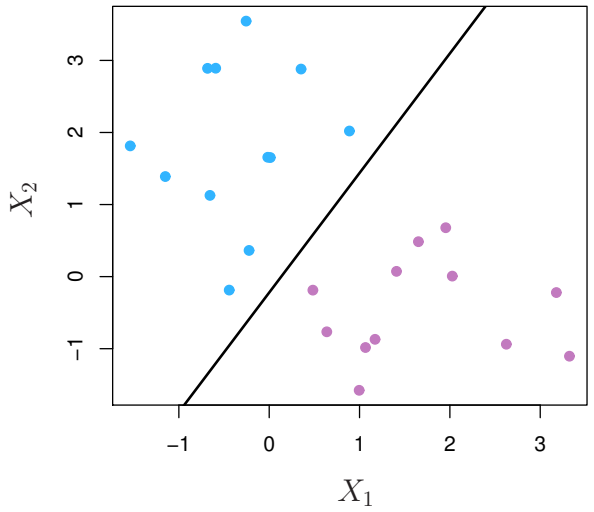


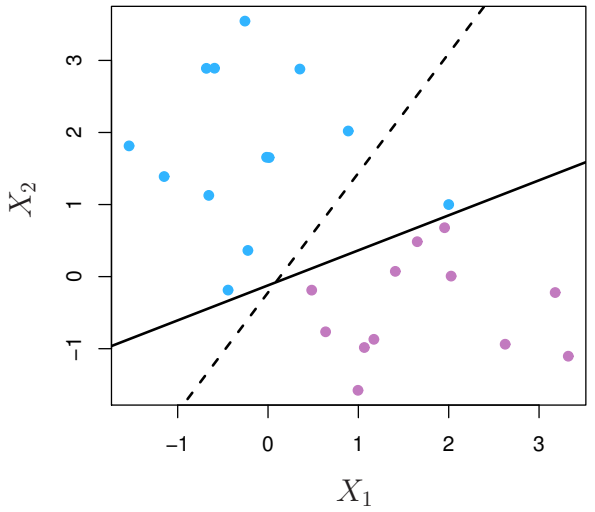
Maximal Margin Classifier

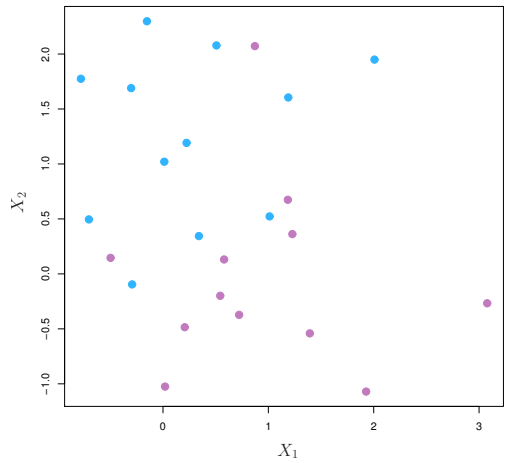






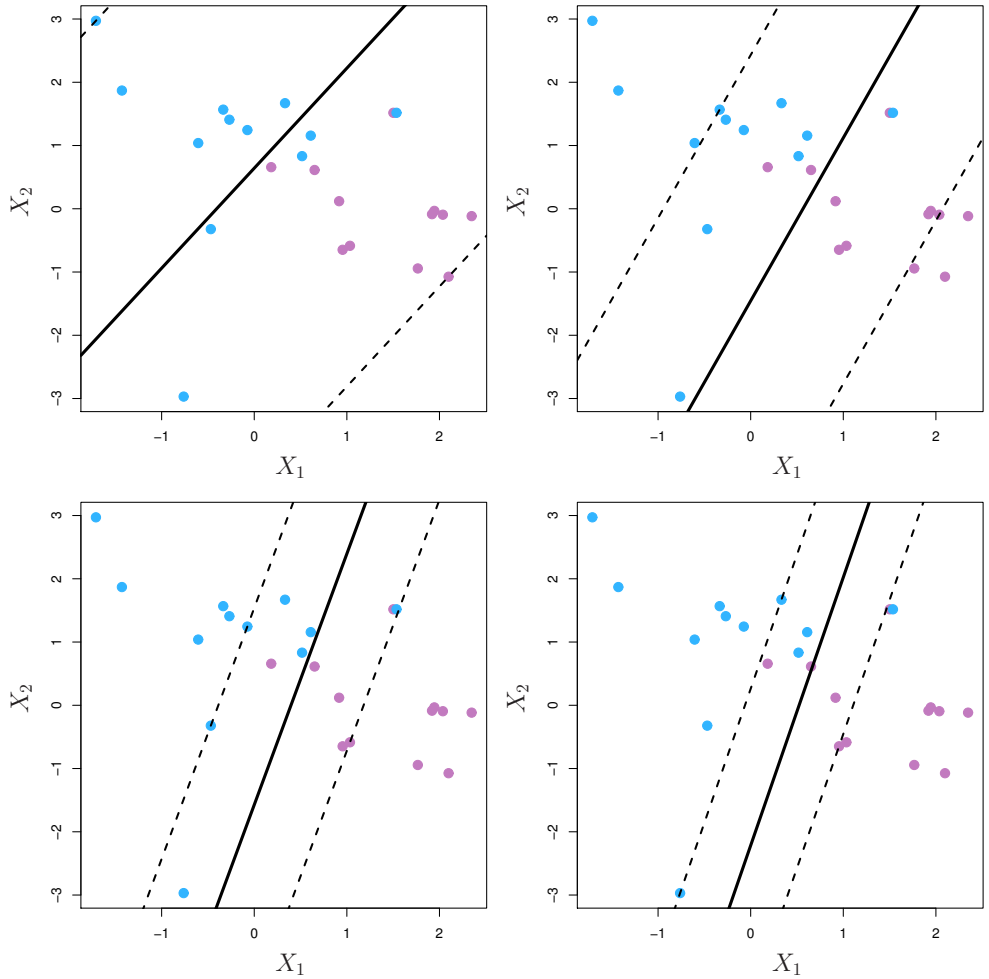






Support Vector Classifier

We allow “few” players from the other team to be on our side (but inside the margin)



Support Vector Classifier

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

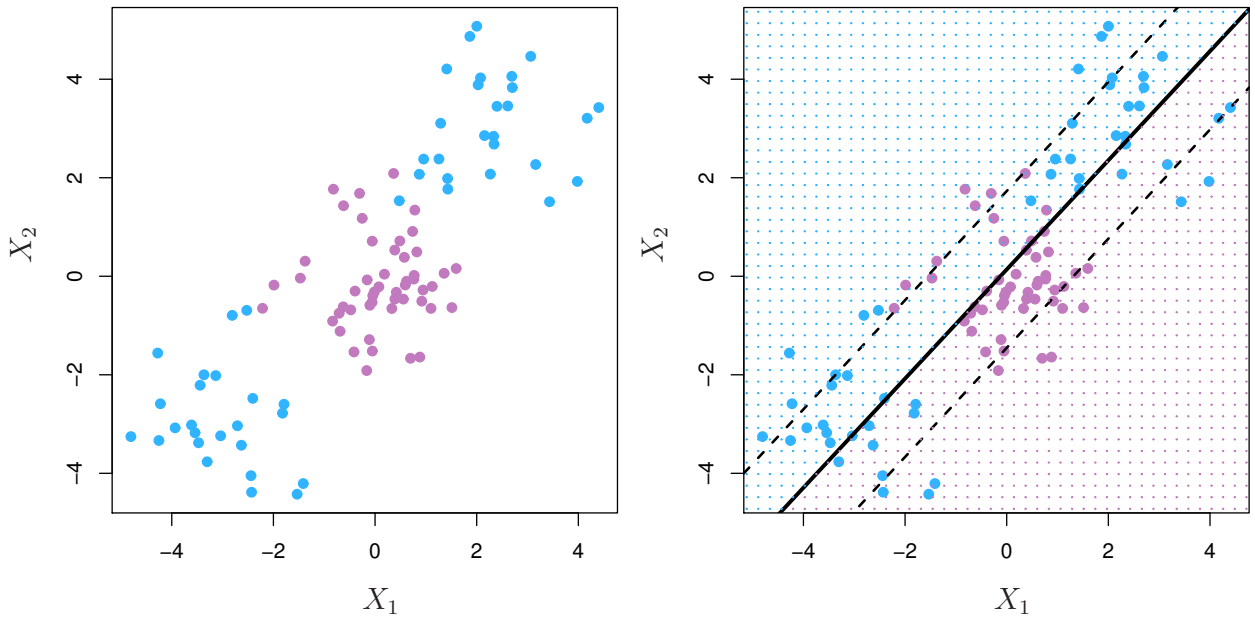
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

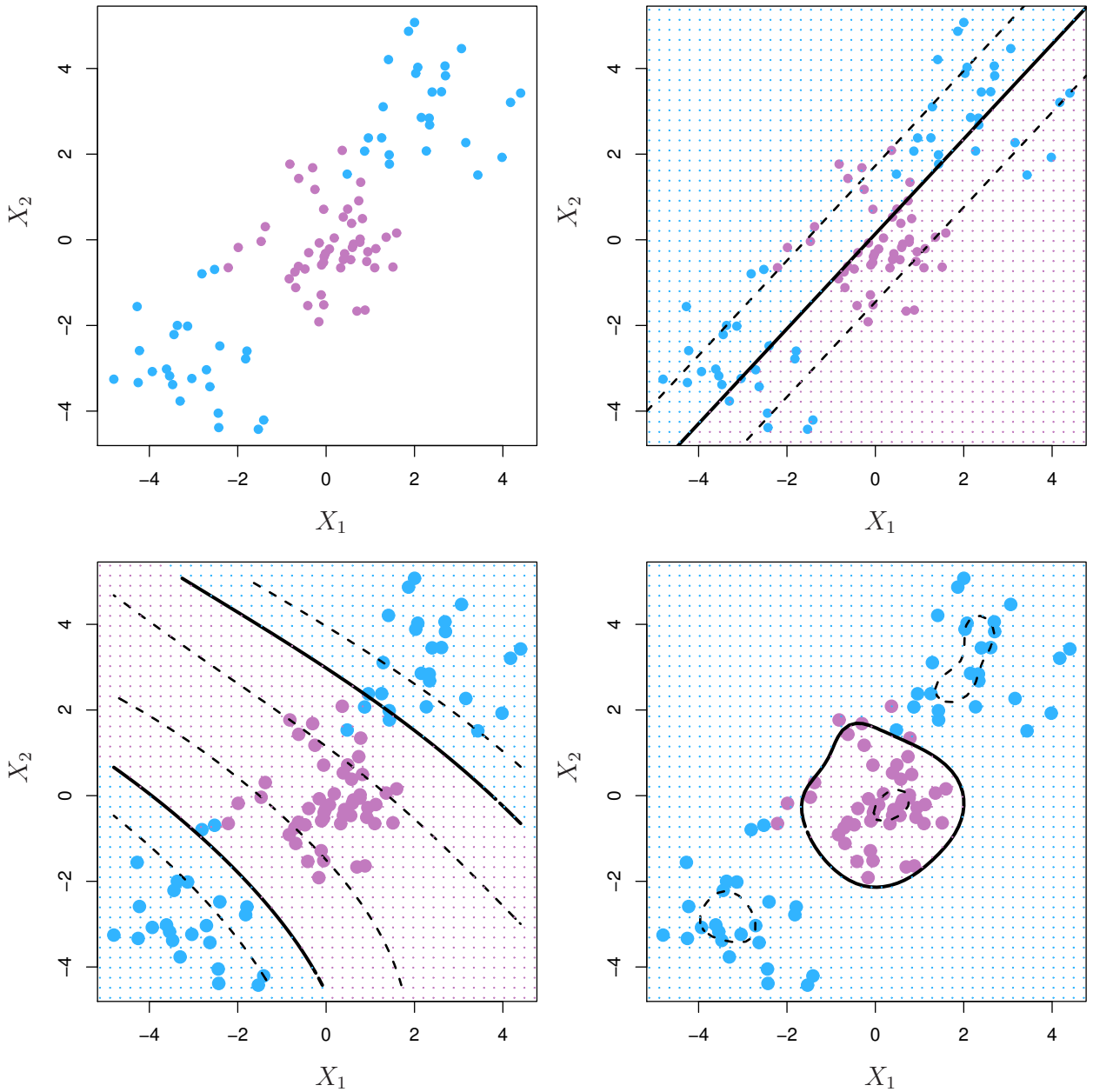
$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

- M : the width of the margin
- C : a nonegative tuning parameter
- ϵ_i: *slack variables* that allow individual observations to be on the wrong side of the margin or the hyperplane ($\epsilon_i = 0, 0 < \epsilon_i < 1, \epsilon_i > 1$).

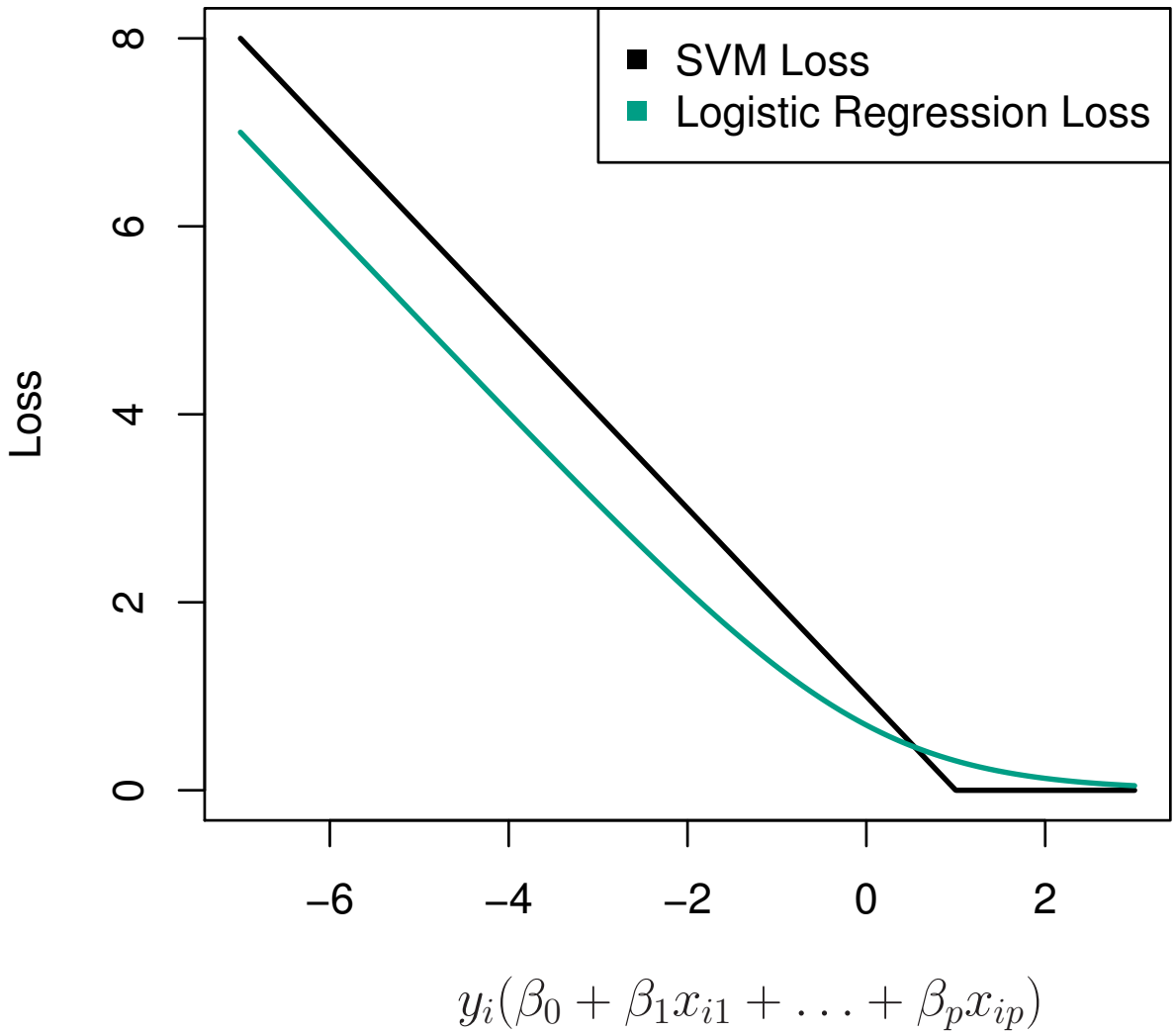
Support Vector Machines



Support Vector Machines



SVM and logistic regression



Classification methods

- Logistic regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- k-Nearest Neighbours
- Support Vector Machines
- Advanced methods: Trees, Boosting and Random Forests (Week 10)

$\rightarrow$ Generative and discriminative models

Which classification method?

- Is it 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

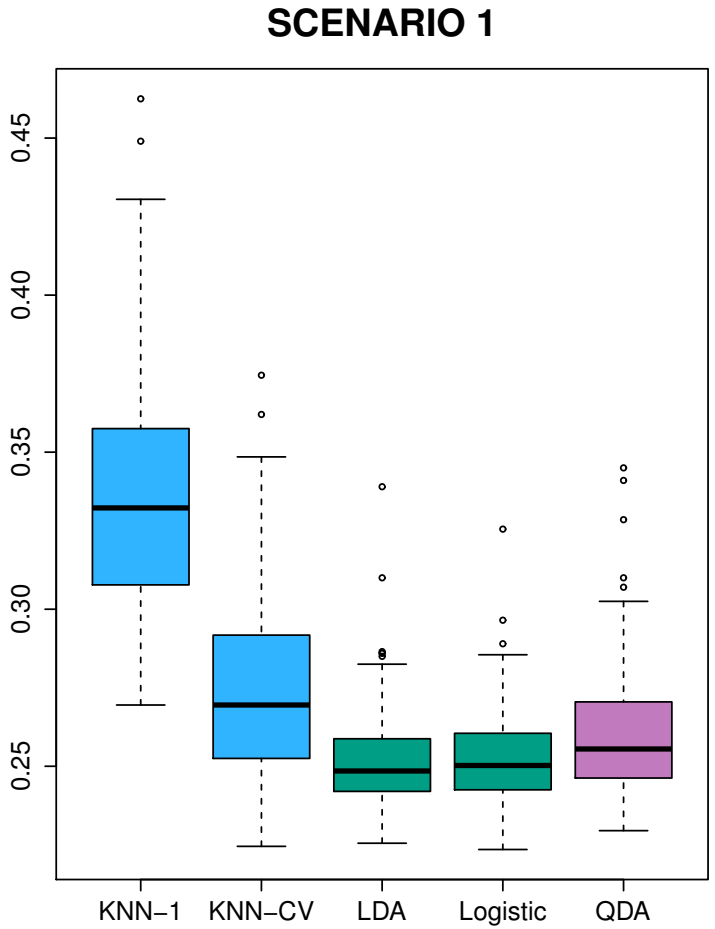
- We compare the following classifiers: KNN-1, KNN-CV, LDA, Logistic and LDA
- We consider **six different scenarios** for the data generating process
- Scenarios 1-3 are **linear**, and scenarios 4-6 are **nonlinear**
- In each scenario, we generate 100 **random training data sets**. For each of these 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.

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.

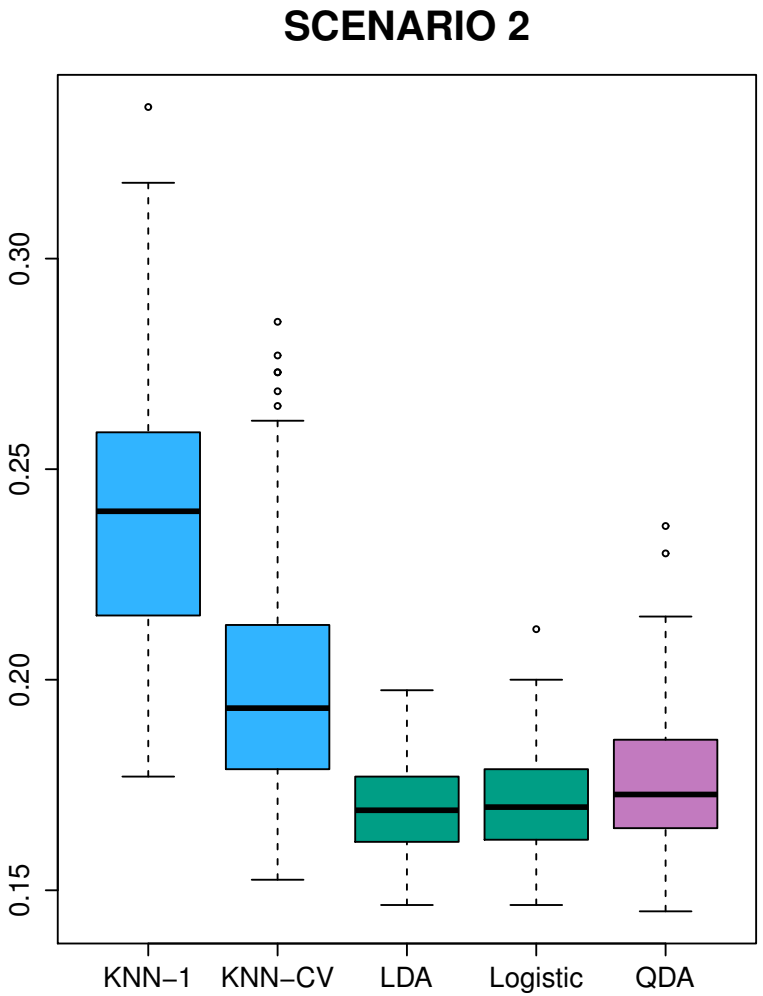


Scenario 2

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

Scenario 2

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

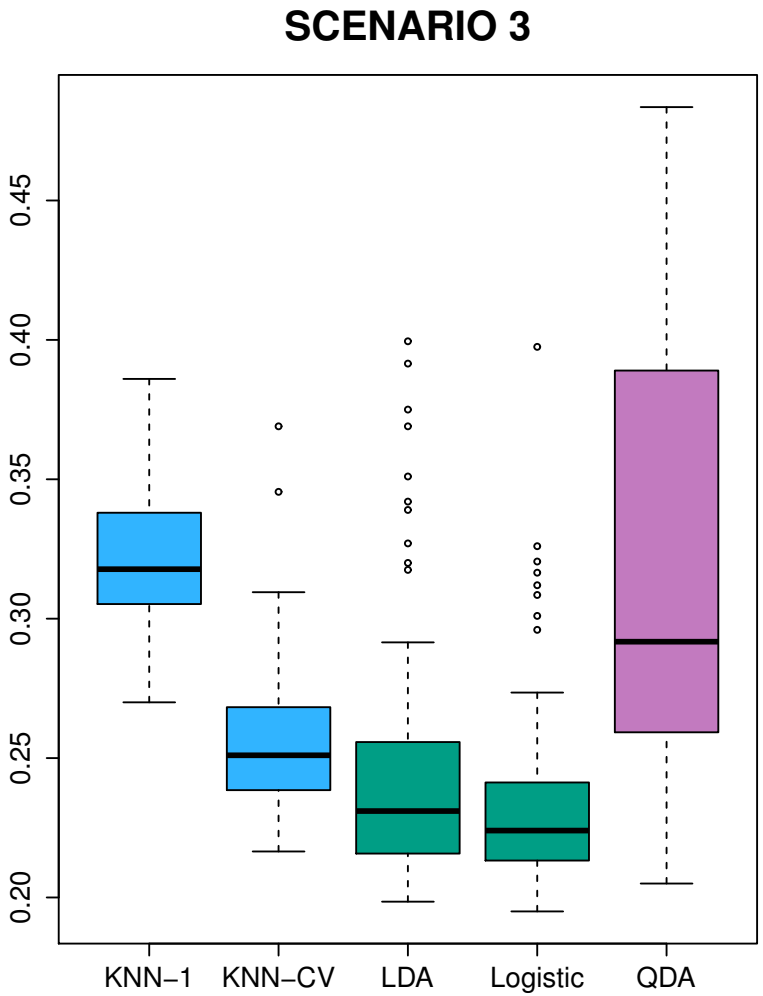


Scenario 3

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

Scenario 3

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

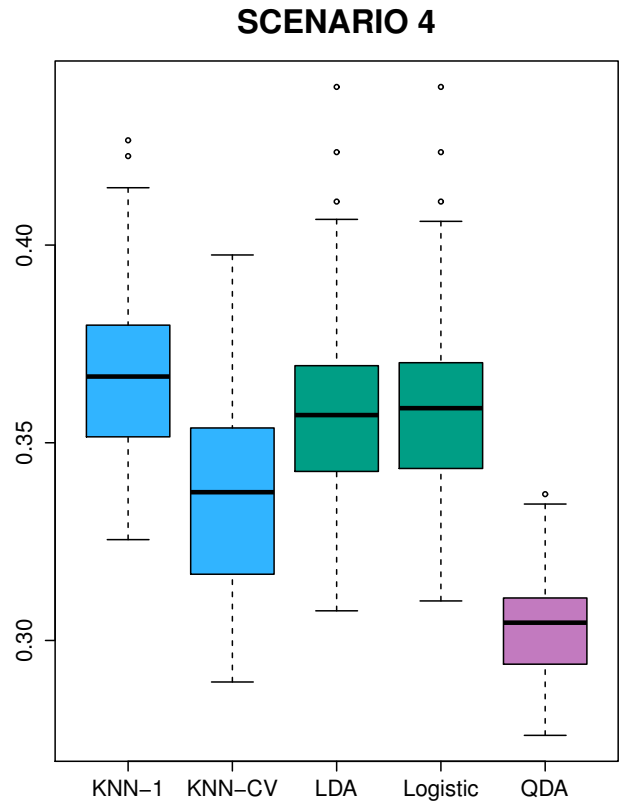


Scenario 4

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.

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.

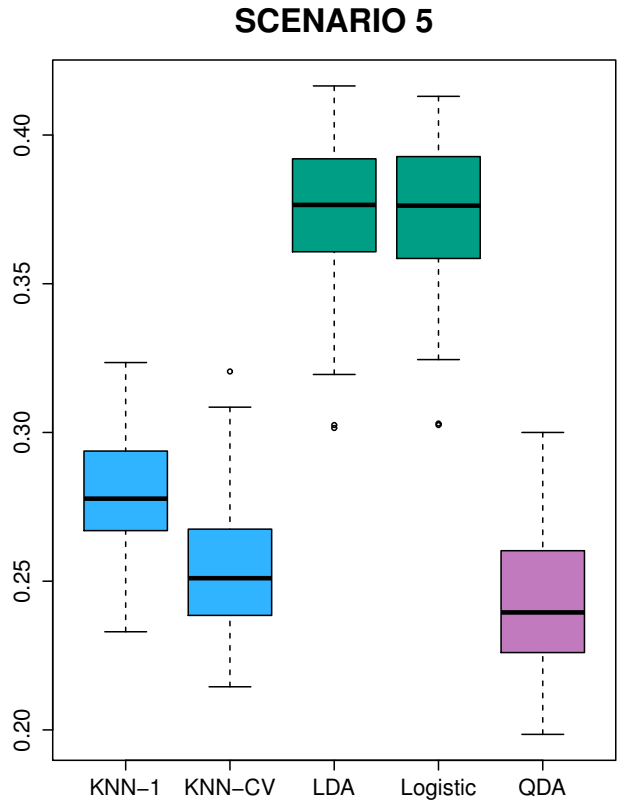


Scenario 5

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.

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.

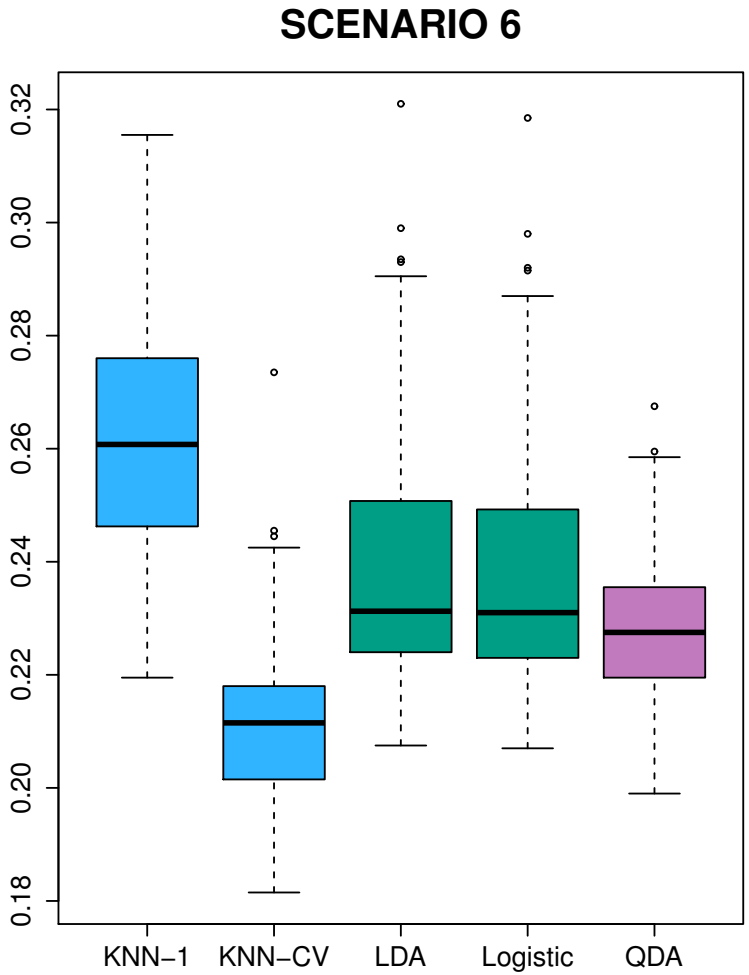


Scenario 6

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

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 (Week 10)