Statistical learning - Classification

Rita Almeida

PhD course 3045, VT 2018

Classification

Regression is used when the outcome variable is quantitative.

Classification is used when the output variable is categorical.

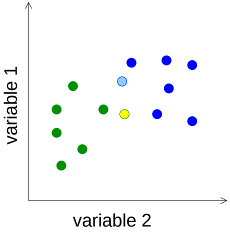
- Inference
- Prediction
  - The aim is to decide to what class C_j a new input x belongs.

K-nearest neighbors

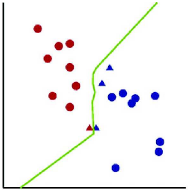
Classification without learning a separation function.
(Use the data for prediction.)

Nearest-neighbor (NN):

Simple



K-nearest neighbors (KNN)



Misaki et al., Neuroimage 2010

For a new x the KNN classifier estimates $p(C_j|x)$ as the fraction of K nearest neighbours of x that belong to C_j.

Classification

Classification learning a separation function.

Discriminative model classifiers: Estimate $p(C_j|x)$ directly.

- Logistic regression

Generative model classifiers: Estimate $p(C_j)$ and $p(x|C_j)$ to calculate $p(C_j|x)$ using Bays theorem.

- Linear discriminant analysis
- Quadratic discriminant analysis

Discriminant classifiers: Find a discriminant function that directly maps an input x to a category C_j.

- Support vector classifiers

Logistic regression

Discriminative model classifiers: Estimate $p(C_j|x)$ directly.

- Logistic regression (2 categories): $p(C_j|x)$ is estimated by a logistic function of a linear combination of the independent variables.

$$Y = \beta_0 + \beta_1 X + \dots \qquad p(X) = \frac{e^{\beta_0 + \beta_1 X + \dots}}{1 + e^{\beta_0 + \beta_1 X + \dots}} \qquad log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X + \dots$$

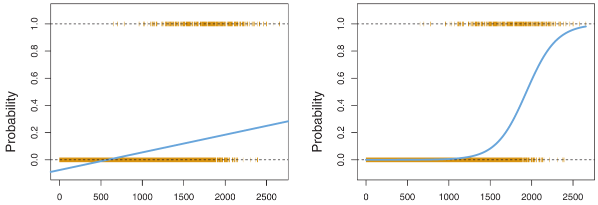


Figure adapted from James et al.

- Multinomial regression (more than 2 categories)

Discriminant analysis

Generative model classifiers: Estimate $p(C_j)$ and $p(x|C_j)$ to calculate $p(C_j|x)$ using Bays theorem.

- Discriminant analysis : assumes that the $p(x|C_j)$ are normally distributed.
  - same covariance matrix for different categories - Linear DA
  - different covariance matrix for different categories - Quadratic DA.

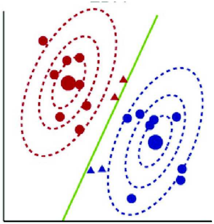


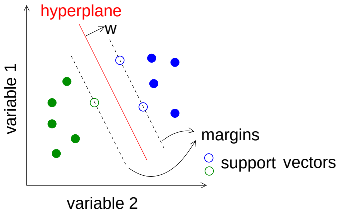
Figure adapted from Misaki et al.,

Neuroimage 2010

Support vector classifiers

Discriminant classifiers: Find a discriminant function (a hyperplane) that directly maps an input x to a category C_j.

- Support vector classifiers: maximize the margin separating two categories.



- All the points on the margin are support vectors.

- The maximal margin hyperplane in p dimensions is defined based on a set of weights $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.
- A new observation x^* is classified based on the sign of $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$

Soft-margin classifiers

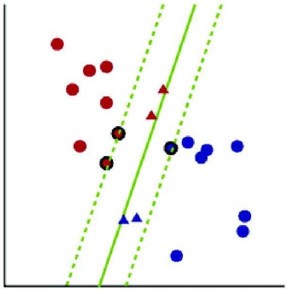


Figure adapted from Misaki et al., Neuroimage 2010

- All the points on the margin or on the wrong side of the margin are support vectors.

Soft-margin classifiers used for:

- non-separable classes,
- greater robustness of the classifier to individual observations.

Support vector classifiers

SVC is the solution to the optimization problem:

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

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

$$y_i(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \geq M(1 - \epsilon_i)$$

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

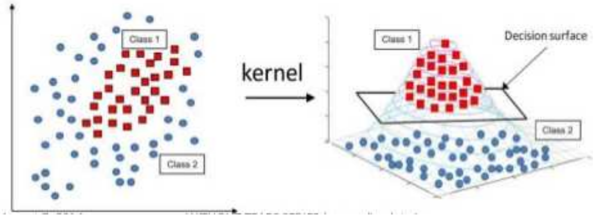
- $\epsilon_1, \ldots, \epsilon_n$ are slack variables.
  - If the observation i is on the wrong side of the margin then $\epsilon_i > 0$.
  - If the observation i is on the wrong side of the hyperplane then $\epsilon_i > 1$.

Support vector classifiers

- C determines the trade-off between model complexity and points between boundaries or misclassified.
  - Small C few points are allowed inside the margins $\rightarrow$ potential low bias but high variance
  - Large C many points are allowed inside the margins or misclassified $\rightarrow$ potential high bias but low variance
- C can be set by cross-validation.
- All the points on the margin or on the wrong side of the margin are support vectors.
  - Large C results in more support vectors.
- Observations that are not support vectors do not affect the classifier.

Non-linear support vector classifiers

- The features can include polynomials or other functions of the original variables.
- The feature space can be increased using non-linear kernels.
- The usage of kernels leads to an efficient approach to find non-linear boundaries between classes.



Support vector classifiers

For more than 2 categories

- Classification using all possible pairs, plus a voting scheme.
- Classification of each category versus all other.

Example: ADHD classification

ADHD-200 Consortium:

- data from 8 imaging centers
- children and adolescents (7-21 years)
- 491 typical developing, 285 diagnosed with ADHD
- resting-state fMRI
- structural MRI
- phenotypic information including: diagnostic status, dimensional ADHD symptom measures, age, sex, intelligence quotient (IQ)

Competition:

- 197 extra datasets (from six sites) without labels
- aim was to give correct diagnosis: typically developing, ADHD primarily inattentive type, or ADHD combined type
- 21 teams submitted solutions
- evaluation: 1 point awarded per correct diagnosis; 1/2 point for correct ADHD with subtype incorrect.

Example: ADHD classification

- A team from University of Albert had highest accuracy using: age, sex, handedness and IQ!

Diagnostic task	Input data	Classifier	Holdout accuracy (%)
Binary	Chance		55.0
	*PCs*1	Quadratic SVM	69.0
	*PCs*2	Linear SVM	65.5
Three-way	Chance		55.0
	*PCs*1	Logistic	63.7
	*PCs*2	Logistic	59.1

Table adapted from Brown et al., Front Syst Neurosci. 2012

Example: ADHD classification

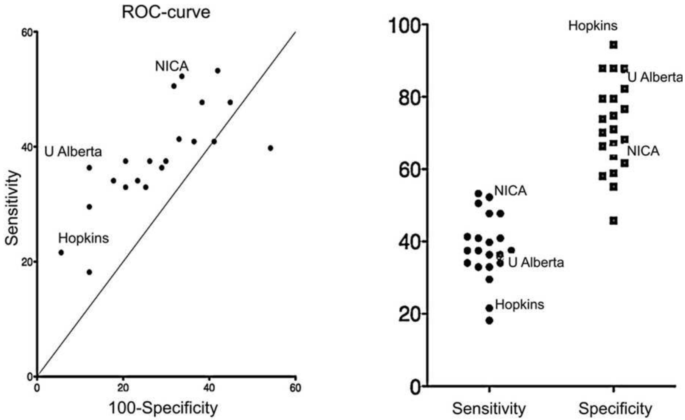


Figure adapted from ADHD-200 Consortium, Front Syst Neurosci. 2012