Introduction to classification

Gianluca Campanella

Contents

Classification

Regression versus classification

Regression

Aim Predict a continuous value

Loss How ‘off’ (numerically) our predictions are

Classification

Aim Predict a class

Loss How ‘inaccurate’ the predicted classes are

k-nearest neighbours classifier

k-nearest neighbours classifier

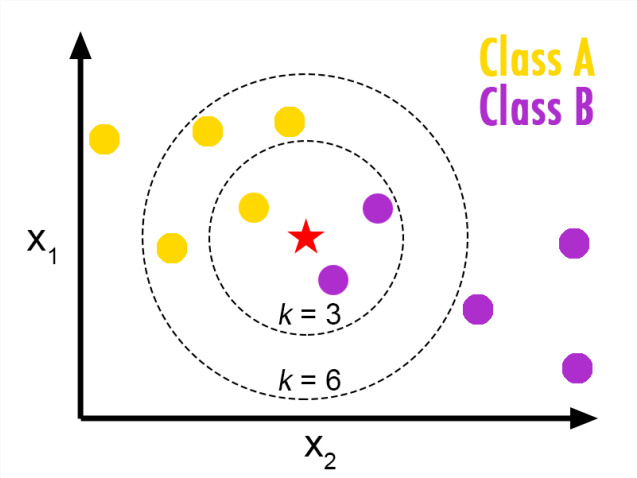
Given a new observation…

- Find the *k* ‘most similar’ training sample(s)
- Use the most common class among them as prediction

Questions

- How do we define similarity?
- How many neighbours do we use?

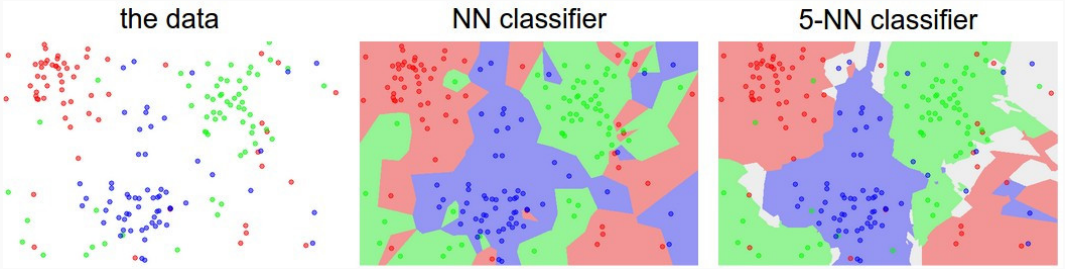
k-nearest neighbours classifier



From Burton DeWilde's blog

Choice of k

- Larger $k \rightarrow$ smoother boundaries, less 'noisy'
- If $k = N$, we always predict the majority class



From *CS231n: Convolutional Neural Networks for Visual Recognition*

Minkowski distance

$$\left(\sum_i |x_i - y_i|^p \right)^{1/p}$$

$p = 1$	Manhattan distance	$\sum_i \lvert x_i - y_i \rvert$
$p = 2$	Euclidean distance	$\sqrt{\sum_i (x_i - y_i)^2}$

Weight of neighbours

Uniform weights

- All k neighbours contribute equally to the prediction
- Actual distance to each is ignored

Distance weights

- Contributions are weighted by $1/\text{distance}$
- Closer neighbours influence the prediction more

Curse of dimensionality

As the number of variables (coordinates) increases…

- The volume of the space increases
- Pairwise distances become more similar $\rightarrow$ sparsity
- Some samples have huge neighbourhoods $\rightarrow$ ‘hubs’

Metrics

Classification accuracy

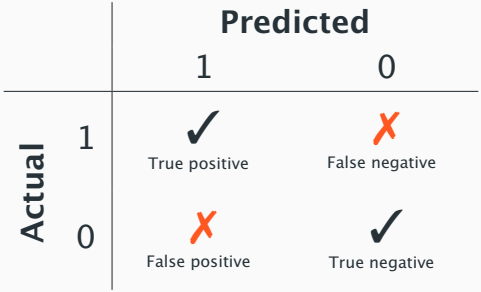
Classification accuracy

- Percentage of correct predictions
- Higher is better

Classification error

- Percentage of incorrect predictions (inverse of accuracy)
- Lower is better

Confusion matrix



- Gives a better understanding of behaviour
- Can be used to define multiple performance metrics

Sensitivity and specificity

Sensitivity
(a.k.a. true positive rate)

$$\frac{\sum \text{True positive}}{\sum \text{Actual} = 1}$$

Specificity
(a.k.a. true negative rate)

$$\frac{\sum \text{True negative}}{\sum \text{Actual} = 0}$$

Sensitivity and specificity

Sensitivity
(a.k.a. true positive rate)

$$\frac{\sum \text{True positive}}{\sum \text{Actual} = 1}$$

Specificity
(a.k.a. true negative rate)

$$\frac{\sum \text{True negative}}{\sum \text{Actual} = 0}$$

Perfect sensitivity

- All sick identified as sick
- Negative test result definitely rules out disease

Sensitivity and specificity

Sensitivity
(a.k.a. true positive rate)

$$\frac{\sum \text{True positive}}{\sum \text{Actual} = 1}$$

Perfect sensitivity

- All sick identified as sick
- Negative test result definitely rules out disease

Specificity
(a.k.a. true negative rate)

$$\frac{\sum \text{True negative}}{\sum \text{Actual} = 0}$$

Perfect specificity

- No healthy identified as sick
- Positive test result useful for ruling in disease

Sensitivity and specificity

Sensitivity
(a.k.a. true positive rate)

$$\frac{\sum \text{True positive}}{\sum \text{Actual} = 1}$$

Specificity
(a.k.a. true negative rate)

$$\frac{\sum \text{True negative}}{\sum \text{Actual} = 0}$$

Perfect sensitivity

- All sick identified as sick
- Negative test result definitely rules out disease

Perfect specificity

- No healthy identified as sick
- Positive test result useful for ruling in disease

Can we maximise both at the same time?

‘Perfect’ border control

100% sensitivity

- ‘Everyone is a terrorist!’

‘Perfect’ border control

100% sensitivity

- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity

‘Perfect’ border control

100% sensitivity

- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity
- No one can enter the country! $\rightarrow$ 0% specificity

‘Perfect’ border control

100% sensitivity

- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity
- No one can enter the country! $\rightarrow$ 0% specificity

100% specificity

- ‘No one is a terrorist!’

‘Perfect’ border control

100% sensitivity

- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity
- No one can enter the country! $\rightarrow$ 0% specificity

100% specificity

- ‘No one is a terrorist!’
- All non-terrorists are allowed in $\rightarrow$ 100% specificity

‘Perfect’ border control

100% sensitivity

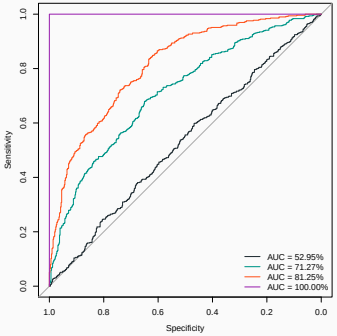
- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity
- No one can enter the country! $\rightarrow$ 0% specificity

100% specificity

- ‘No one is a terrorist!’
- All non-terrorists are allowed in $\rightarrow$ 100% specificity
- All terrorists are also allowed into the country! $\rightarrow$ 0% sensitivity

ROC and AUC

Receiver Operating Characteristic (ROC) curve

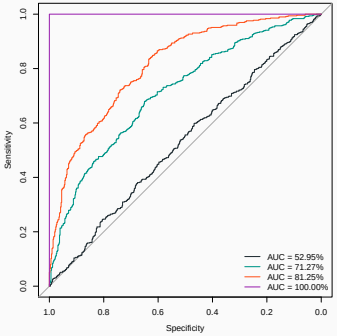


Sensitivity *vs* (1 − specificity)

$\rightarrow$ TP rate *vs* FP rate

ROC and AUC

Receiver Operating Characteristic (ROC) curve



Sensitivity *vs* (1 − specificity)

$\rightarrow$ TP rate *vs* FP rate

Area Under the Curve (AUC)

- Probability that Prediction(actual 1) > Prediction(actual 0)
- Random guess $\rightarrow$ AUC = 50% (diagonal)
- Higher is better

Cost-benefit analysis

Cost-benefit analysis

- Assume that the four possible outcomes of a classification problem have (numerical) benefits and costs
  - > 0 desirable (e.g. profit)
  - $= 0$ neutral
  - < 0 undesirable (e.g. loss)
- These ‘benefits and costs’ needn’t be symmetrical

Planning an outdoor activity

You have 20 people enrolled in an outdoor activity paying £30 each

- Before the activity, you check the weather forecast and either:
  - Go ahead, which costs you £5 per participant
  - Cancel and refund the participants in full
- If you decide to go ahead, the day of the activity it will either:
  - Be sunny, in which case you get to keep the profit
  - Rain, in which case you'll have to refund the participants in full

What is the 'benefits and costs' matrix?

Planning an outdoor activity

Free weather forecast

Actual \ Forecast	☀	⛈
☀	20%	20%
⛈	10%	50%

Weather forecast costing £15

Actual \ Forecast	☀	⛈
☀	30%	5%
⛈	10%	55%

- What are the sensitivity and specificity?
- What is the expected profit?