An introduction to classification

Dr Gianluca Campanella

26th May 2016

Contents

Classification

Regression versus classification

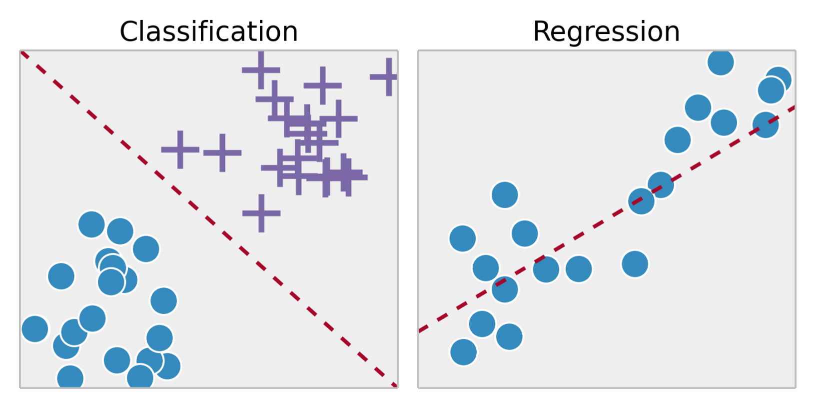
Regression

- Predict a **continuous** value
- Minimise a **loss function** that measures how ‘off’ (numerically) our predictions are

Classification

- Predict a **class**
- Minimise a **loss function** that measures how ‘inaccurate’ the predicted classes are

Regression versus classification



k-nearest neighbours classifier

k-nearest neighbours classifier

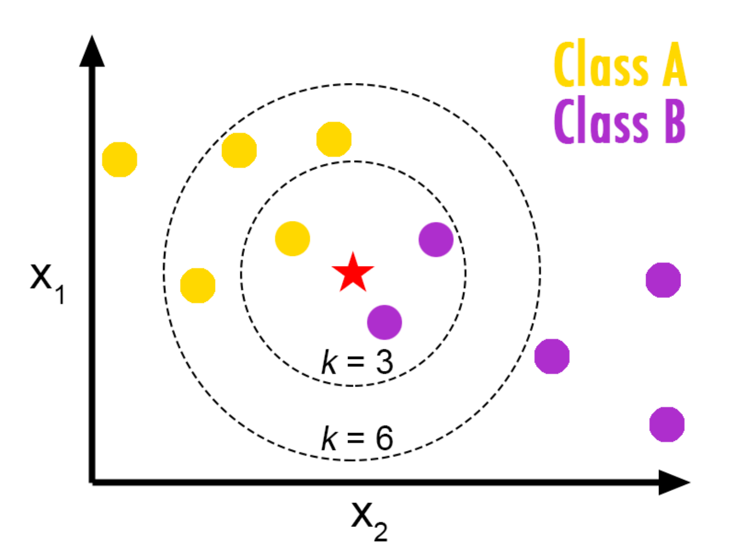
Given a new observation...

- Find the k ‘most similar’ training sample(s)
- Predict the most common class among them

Questions

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

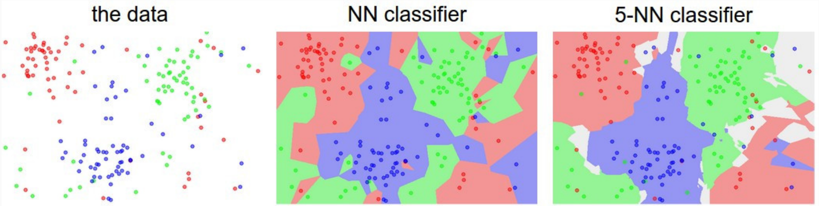
k-nearest neighbours classifier



k-nearest neighbours classifier

Choice of k

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



Distance metrics

Euclidean distance

$$\sqrt{\sum_i (x_i - y_i)^2}$$

`KNeighborsClassifier(…, metric=‘euclidean’, …)`

Manhattan distance

$$\sum_i |x_i - y_i|$$

`KNeighborsClassifier(…, metric=‘manhattan’, …)`

Distance metrics

Minkowski distance

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

```
KNeighborsClassifier(…, metric='minkowski', …)
```

- $p = 1 \rightarrow$ Manhattan distance
- $p = 2 \rightarrow$ Euclidean distance

Weight of neighbours

Uniform weights

- All k neighbours contribute equally to the prediction
- Actual distance to each is ignored
- `KNeighborsClassifier(…, weights='uniform', …)`

Distance weights

- Contributions are weighted by 1/distance
- Closer neighbours influence the prediction more
- `KNeighborsClassifier(…, weights='distance', …)`

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 (inverse of accuracy)

- Percentage of **incorrect** predictions
- Lower is better

Confusion matrix

		Predicted	
		1	0
Actual	1	✓ True positive	✗ False negative
	0	✗ False positive	✓ True negative

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

Sensitivity and specificity

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

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

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 or recall)

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

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

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

Perfect sensitivity:

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

Sensitivity and specificity

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

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

Perfect sensitivity:

- All sick are 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 are identified as sick
- Positive test result useful for ruling in disease

Sensitivity and specificity

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

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

Perfect sensitivity:

- All sick are 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 are identified as sick
- Positive test result useful for ruling in disease

Can we maximise both at the same time?

Example: 'perfect' border control

100% sensitivity

- 'Everyone is a terrorist!'

Example: ‘perfect’ border control

100% sensitivity

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

Example: ‘perfect’ border control

100% sensitivity

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

Example: 'perfect' border control

100% sensitivity

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

100% specificity

- 'No one is a terrorist!'

EXAMPLE: 'perfect' border control

100% sensitivity

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

100% specificity

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

EXAMPLE: ‘perfect’ border control

100% sensitivity

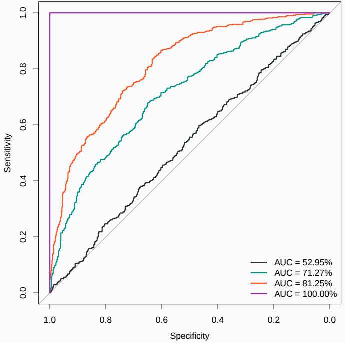
- ‘Everyone is a terrorist!’
- All terrorists are stopped $\rightarrow$ 100% sensitivity
- No one can enter the country!

100% specificity

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

ROC and AUC

Receiver Operating Characteristic (ROC) curve

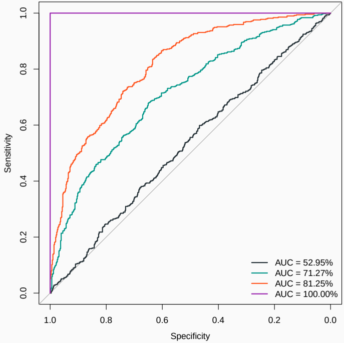


Sensitivity *vs* (1 − specificity)

→ TP rate *vs* FP rate

ROC and AUC

Receiver Operating Characteristic (ROC) curve



Sensitivity *vs* (1 – specificity)
→ TP rate *vs* FP rate

Area Under the Curve (AUC)

- Probability of Prediction(actual 1) > Prediction(actual 0)
- Random guess → AUC = 50% (diagonal)
- Higher is better
- Can be used for **model selection**

Precision and recall

Precision

(a.k.a. positive predictive ratio)

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

Recall

(a.k.a. true positive rate or sensitivity)

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

Precision and recall

Precision

(a.k.a. positive predictive ratio)

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

- How ‘useful’ the search results are
- Perfect precision: only relevant results

Recall

(a.k.a. true positive rate or sensitivity)

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

Precision and recall

Precision
(a.k.a. positive predictive ratio)

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

- How 'useful' the search results are
- Perfect precision: only relevant results

Recall
(a.k.a. true positive rate or sensitivity)

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

- How 'complete' the search results are
- Perfect recall: all relevant results

Precision and recall

Precision

(a.k.a. positive predictive ratio)

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

- How 'useful' the search results are
- Perfect precision: only relevant results

Recall

(a.k.a. true positive rate or sensitivity)

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

- How 'complete' the search results are
- Perfect recall: all relevant results

Can be summarised into an F_1 or F_β score

Cost-benefit analysis

- Assume that the four possible outcomes of the classification have **costs and benefits**
- In economics, this is the '**utility** function'

$$\begin{aligned} &> 0 \quad \text{desirable} \\ &= 0 \quad \text{neutral} \\ &< 0 \quad \text{undesirable} \end{aligned}$$

- Utilities needn't be symmetrical

Exercise: planning an outdoor activity

- You have 20 people enrolled in an outdoor activity costing £30 per participant
- The day before the activity, you check the weather forecast and decide to either:
  - If sunny, go ahead, which costs you £5 per participant
  - If rainy, cancel and refund the participants
- 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 refund the participants

Question: what is the utility (profit) matrix?

Exercise: planning an outdoor activity

The first weather forecast you consider is free, and has the following confusion matrix:

		Forecast ☼	Forecast ////
Actual	☼	20%	20%
Actual	////	10%	50%

Questions

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

Exercise: planning an outdoor activity

The second weather forecast costs £15, and has the following confusion matrix:

		Forecast	
		☼	////
Actual	☼	30%	5%
	////	10%	55%

Questions

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