



TutorialSlide3

University of Glasgow

STATS5099: Data Mining

k-nearest neighbours and linear discriminant analysis

Xiaochen Yang
xiaochen.yang@glasgow.ac.uk

University of Glasgow

This week's content

- Classification: evaluation measures, data splitting
- k-nearest neighbours
- linear discriminant analysis

3/14

University of Glasgow

Model evaluation: scalar measures

- correct classification rate, error rate
- class specific correct classification rate, error rate
- sensitivity, specificity, etc.
- ROC curve and AUC

Remark: provide both the number and **comments**

4/14

University of Glasgow

Model evaluation: sensitivity and specificity

		Predicted class		
		Positive	Negative	Measures
True class	Positive	True positive TP	False negative FN	Sensitivity (TPR) $\frac{TP}{TP+FN}$
	Negative	False positive FP	True negative TN	Specificity (TNR) $\frac{TN}{TN+FP}$
Measures		Positive pred. rate $\frac{TP}{TP+FP}$	Negative pred. rate $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{n}$
		False discovery rate 1-positive pred. rate		

5/14

University of Glasgow

Model evaluation: ROC curve and AUC

By chose different threshold, get points in curve

ROC Curve = 100%
ROC Curve = 75%
ROC Curve = 50%
ROC Curve = 0%

TPR (Sensitivity) vs FPR (1 - Specificity)

Legend: True Positive, True Negative, False Positive, False Negative

Area Under the Curve (AUC) = 1

6/14

University of Glasgow

Data splitting

- Training, validation and test data
- Cross validation
 - K-fold cross validation
 - leave-one-out cross-validation

7/14

University of Glasgow

Data splitting

- Training, validation and test data
- Cross validation
 - K-fold cross validation ($k=10$)
 - leave-one-out cross-validation

When shall we use each method? Comment on advantages and disadvantages.

7/14

University of Glasgow

k-nearest neighbours

k-nearest neighbours (kNN):

- Decide on the value of k ;
- Calculate the distance between the query sample and all the training samples;
- Sort the distances and determine nearest neighbours;
- Gather the categories of the nearest neighbours;
- Use the simple majority rule.

8/14

University of Glasgow

kNN: Task 3

Classify a new paper tissue with acid durability of 3 seconds and strength of 7 kg/square meter.

Acid durability	Strength	Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

9/14

University of Glasgow

kNN: Task 3

Classify a new paper tissue with acid durability of 3 seconds and strength of 7 kg/square meter.

Acid durability	Strength	Distance	Rank	In (3-nn)	Classification
7	7	16	3	Y	Bad
7	4	25	4	N	-
3	4	9	1	Y	Good
1	4	13	2	Y	Good

9/14

University of Glasgow

kNN: conceptual question

Suppose that we take a dataset, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use linear discriminant analysis and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbour and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

for 1-NN
train 20%
test 30%
always be the label
prefer LDA

10/14

University of Glasgow

kNN: summary

Pros:

- Nonparametric (no assumptions on model/data)

Cons:

- High test-time computational cost
- Break down in high dimensional space! (if high dim, do not use knn)

Important considerations:

- Value of k (balance between under- and over-fitting) $k=1$, overfitting, noisy
- Choice of distances (based on data) $k=n$, linear boundary

NB: The list is not exhaustive. (Cont. - Euclidean category - jaccard)

¹Elements of Statistical Learning, §2.5

11/14

University of Glasgow

Linear discriminant analysis (LDA)

- Start with $\Pr(X = x|G = g)$ and assume:
 - Gaussian distribution: for x in group g , $x \sim \mathcal{N}(\mu_g, \Sigma_g)$
 - Equal covariance: $\Sigma_1 = \Sigma_2 = \dots = \Sigma_G = \Sigma$ (linear decision boundary)
- Apply the Bayes' theorem
 $\Pr(G = g|X = x) = \frac{\Pr(X = x|G = g)\Pr(G = g)}{\sum_{r=1}^G \Pr(X = x|G = r)\Pr(G = r)}$
if high $\Rightarrow x \in g$
- Bayes discriminant rule: assign x to the class with the largest linear discriminant function
$$LDF_g(x) = x^T \Sigma^{-1} \mu_g - \frac{1}{2} \mu_g^T \Sigma^{-1} \mu_g + \log \pi_g$$

12/14

University of Glasgow

Linear discriminant analysis

- LDA is a parametric method.
- The distribution assumption is more restrictive but allows group membership probabilities to be calculated.
- Assumption of equal group covariances means linear boundaries in original variable space (bias-variance tradeoff).

quad \Rightarrow more para \Rightarrow bias \downarrow
boundary var \uparrow

13/14

University of Glasgow

Linear discriminant analysis

- LDA is a parametric method.
- The distribution assumption is more restrictive but allows group membership probabilities to be calculated.
- Assumption of equal group covariances means linear boundaries in original variable space (bias-variance tradeoff).

What do you think about the test-time computational cost of LDA, compared with kNN?

13/14

University of Glasgow

Summary: Considerations when choosing methods

- parametric vs nonparametric (non-param needs more data (if assumption is parametric))
- linear vs nonlinear decision boundary (data can be separated by a line?)
- model complexity (tend to under- or over-fit?)
- computational cost (training, test) (how long?)

NB: The list will grow as we see more methods.

PCA - reduce dimension
LDA - classification

train, test acc low \Rightarrow underfitting
training high, test low \Rightarrow overfitting

14/14

- Data splitting: when shall we use each method? what are their advantages and disadvantages?
- What do you think about the test-time computational cost of LDA, compared with kNN?
- Apply cross validation to select the optimal number of principal components for PC regression (Week 2, Example 1).
- Perform kNN and LDA on German dataset. In particular,
 - Shall we use all variables? Why?
 - For kNN, shall we use the Euclidean distance on original variables?
 - For LDA, how to interpret the linear discriminant functions?
 - Which evaluation criteria shall we use?