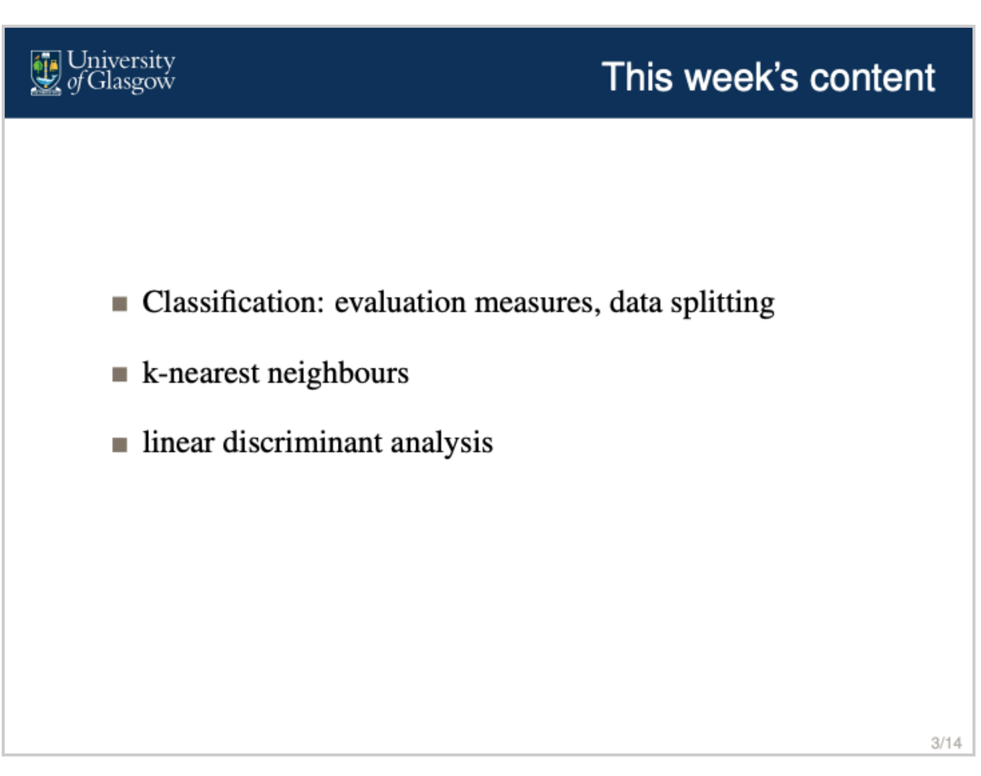
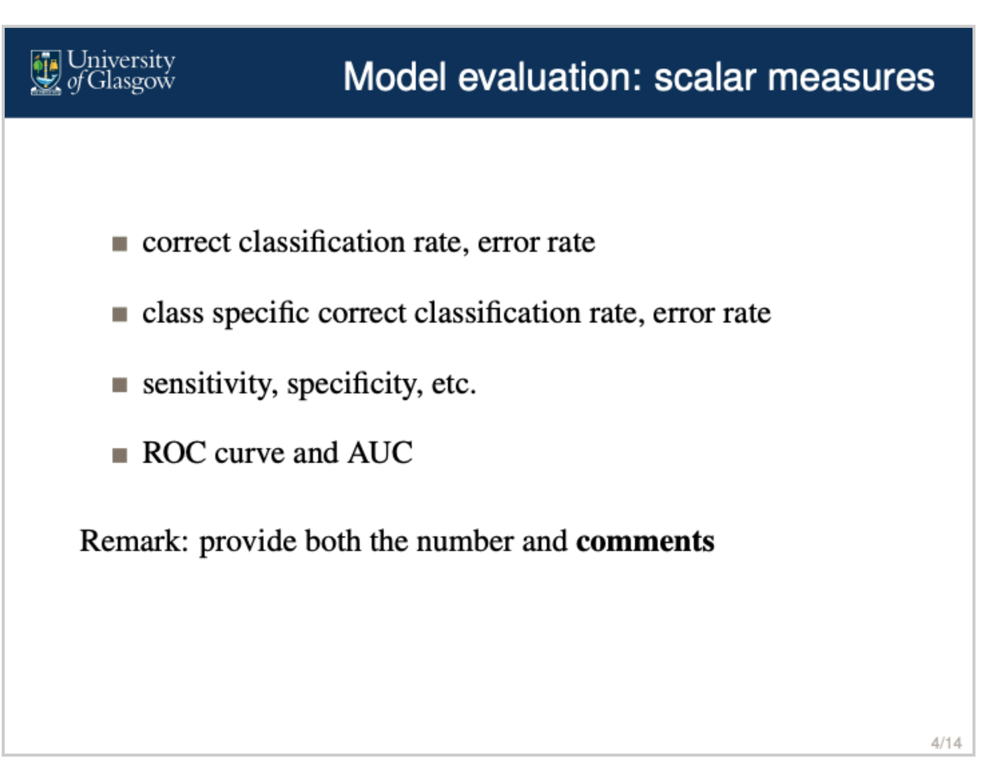
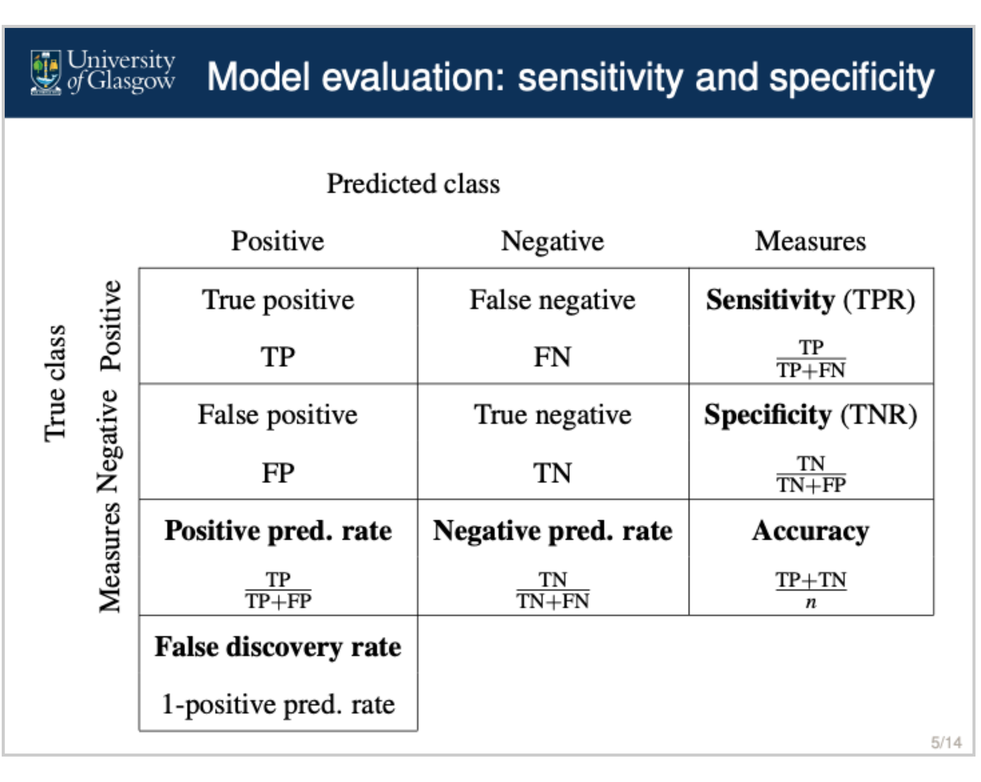
slide3

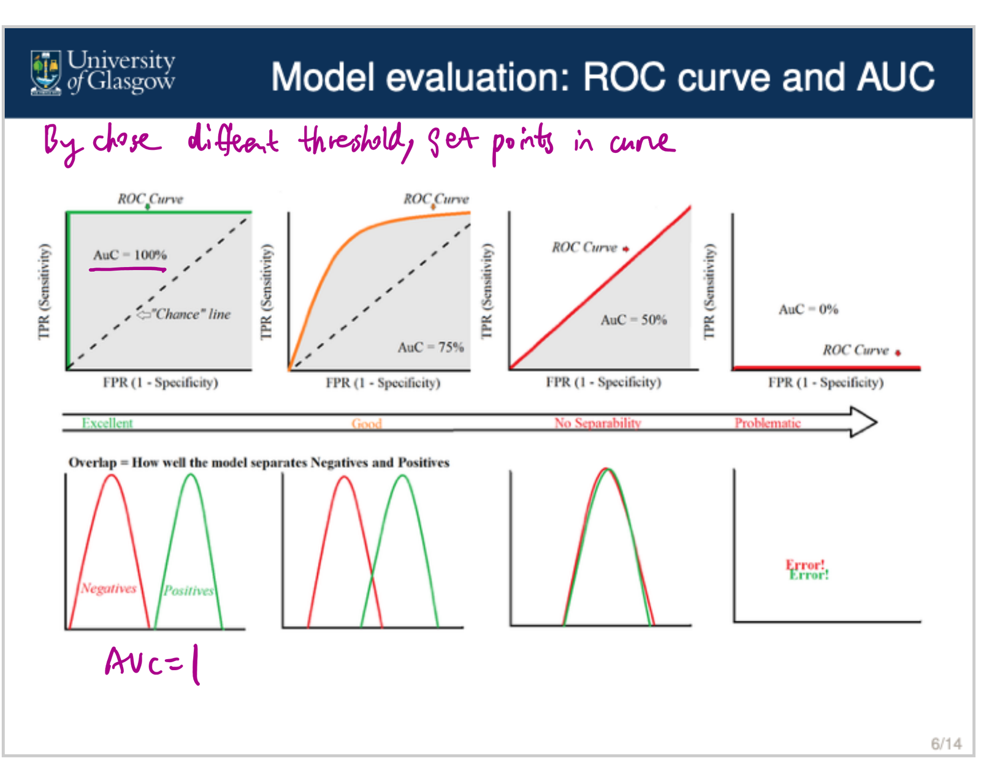
Tuesday, May 25, 2021 4:22 PM

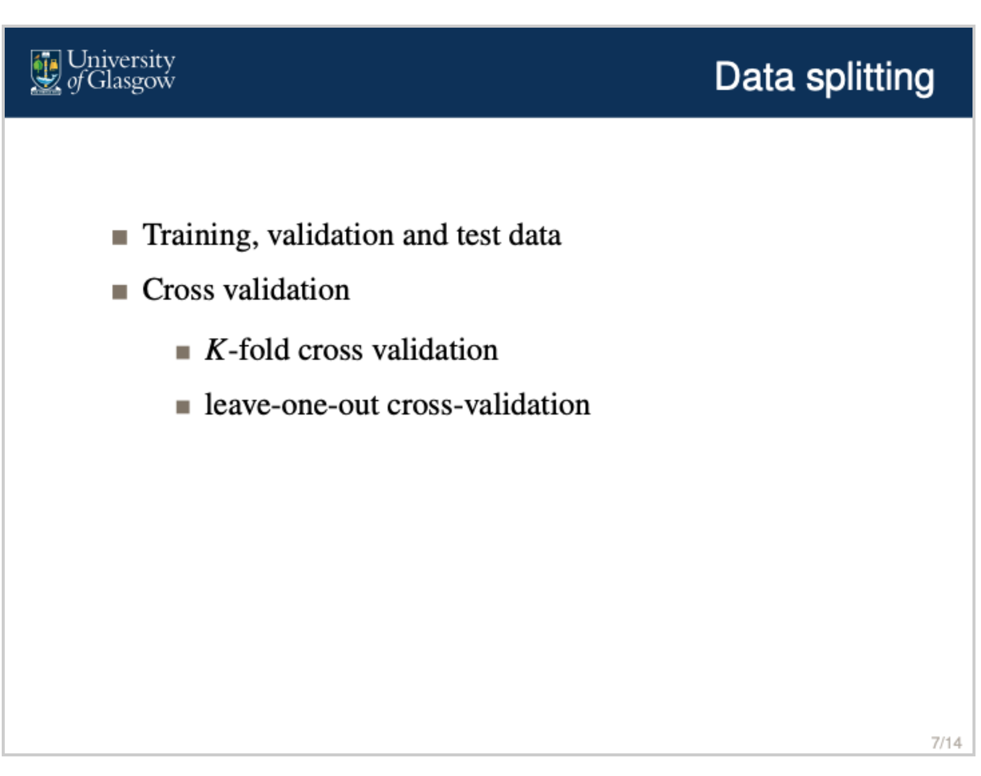
TutorialSlide3

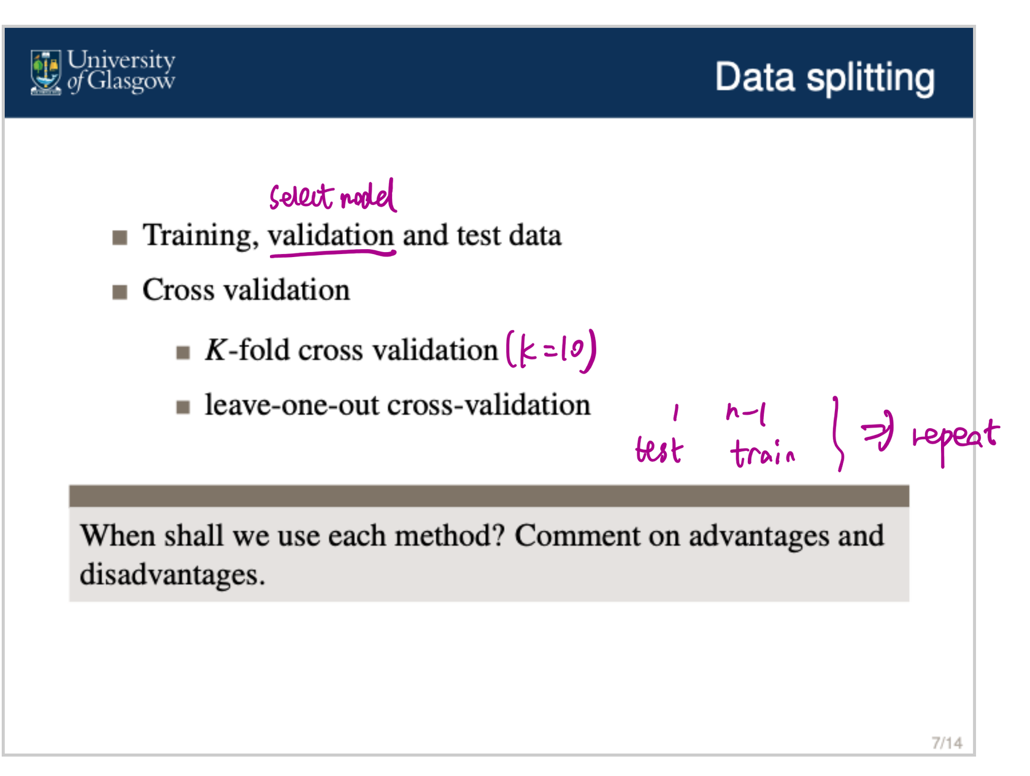


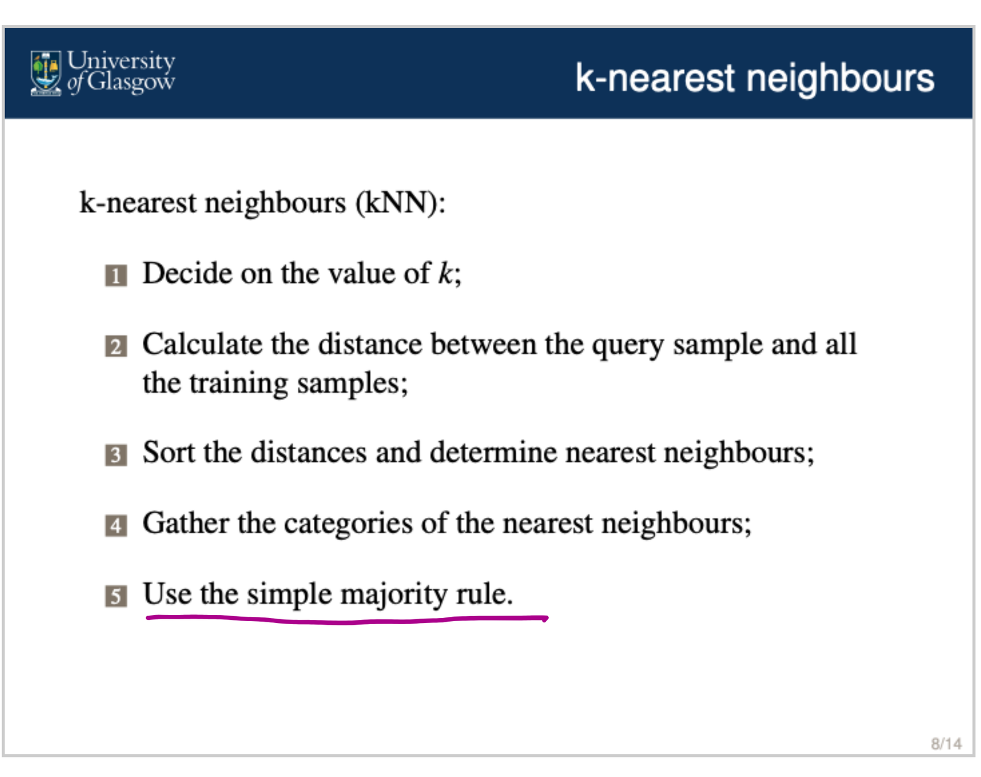


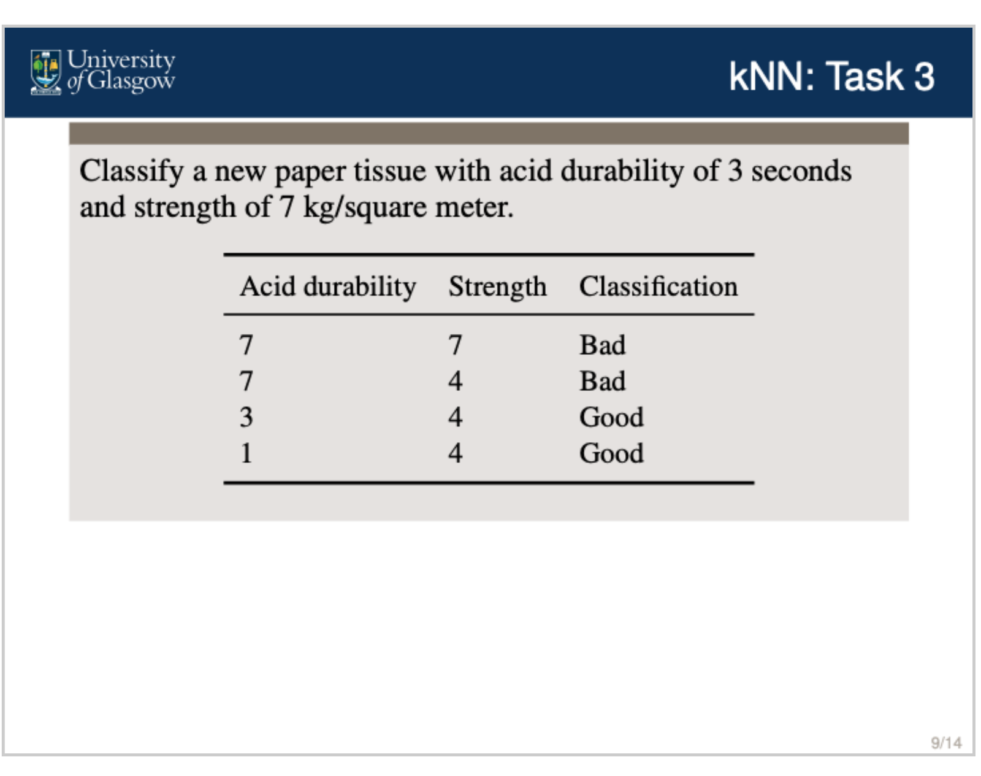

Model evaluation: sensitivity and specificity
Predicted class
Positive
Negative
Measures
True class
Positive
Negative
Measures
True positive TP
False negative FN
Sensitivity (TPR) TP/(TP+FN)
False positive FP
True negative TN
Specificity (TNR) TN/(TN+FP)
Positive pred. rate TP/(TP+FP)
Negative pred. rate TN/(TN+FN)
Accuracy (TP+TN)/n
False discovery rate
1-positive pred. rate
5/14

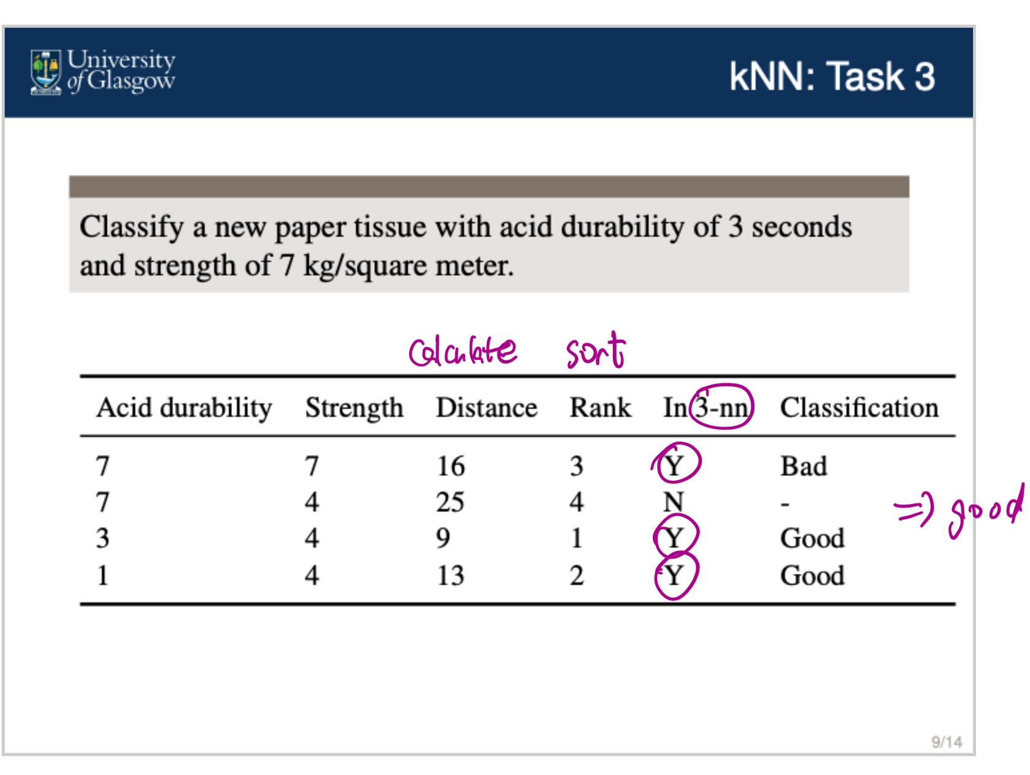


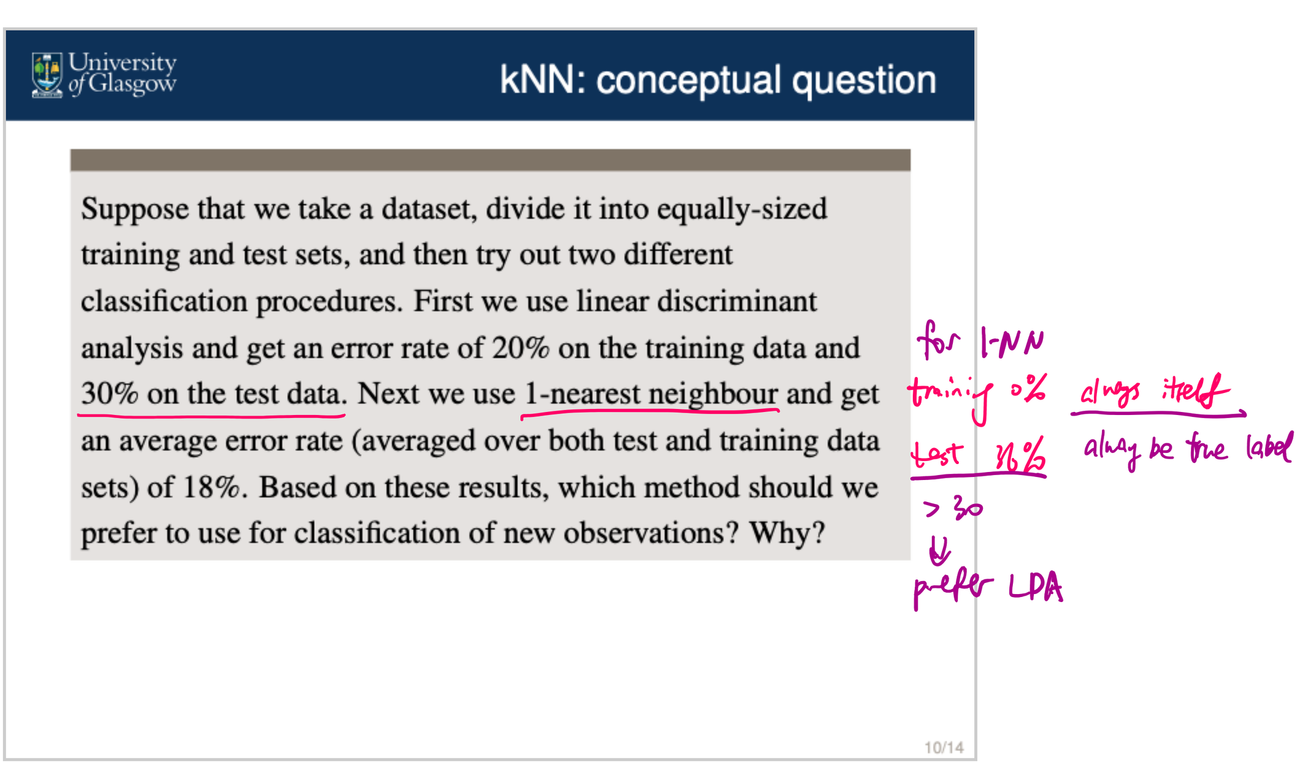


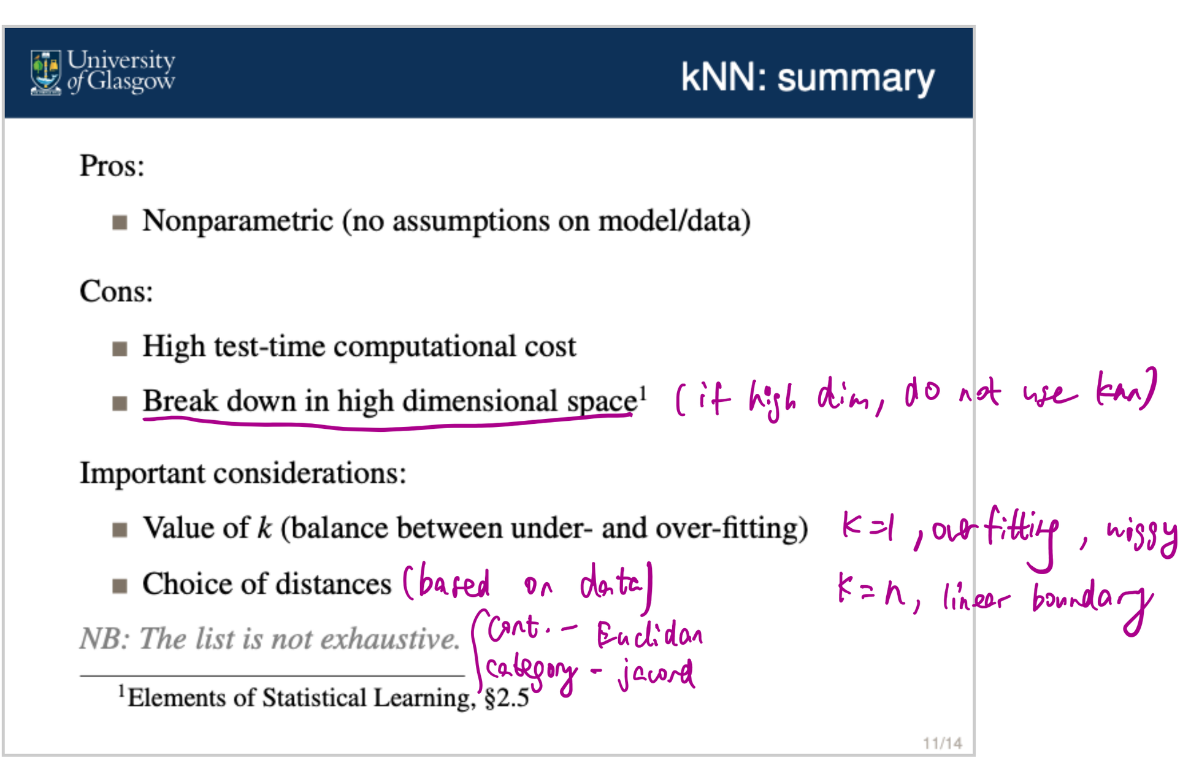


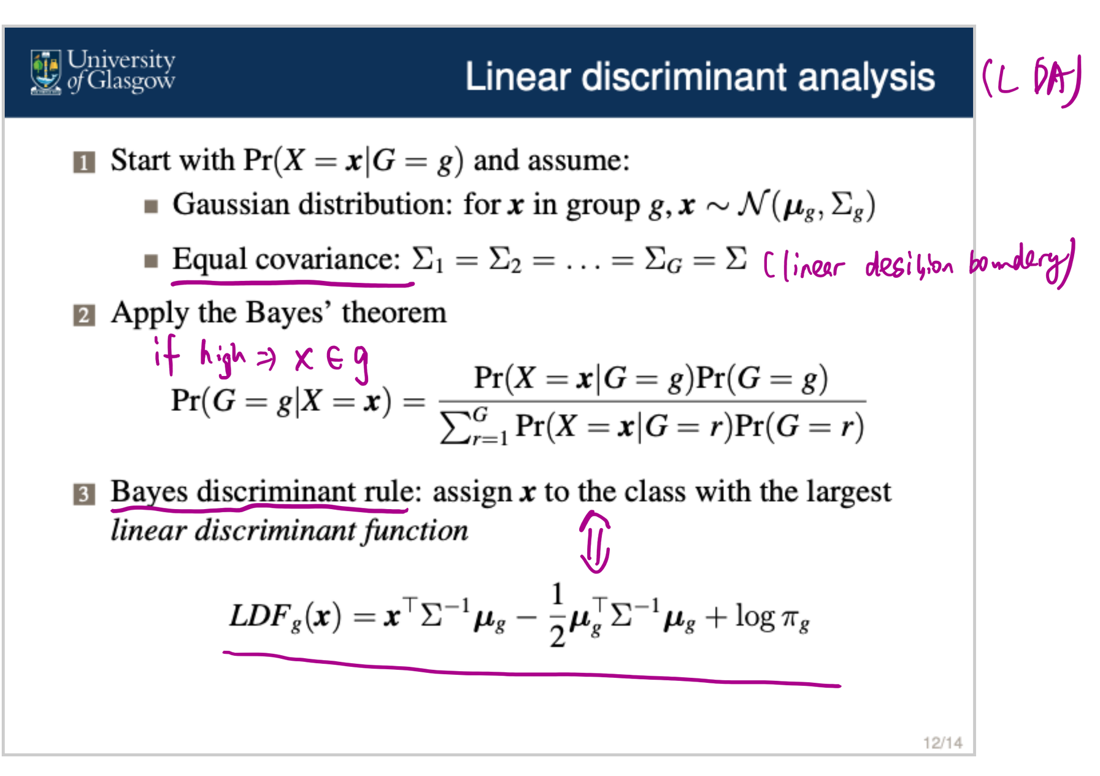

k-nearest neighbours
k-nearest neighbours (kNN):
1 Decide on the value of k;
2 Calculate the distance between the query sample and all the training samples;
3 Sort the distances and determine nearest neighbours;
4 Gather the categories of the nearest neighbours;
5 Use the simple majority rule.
8/14

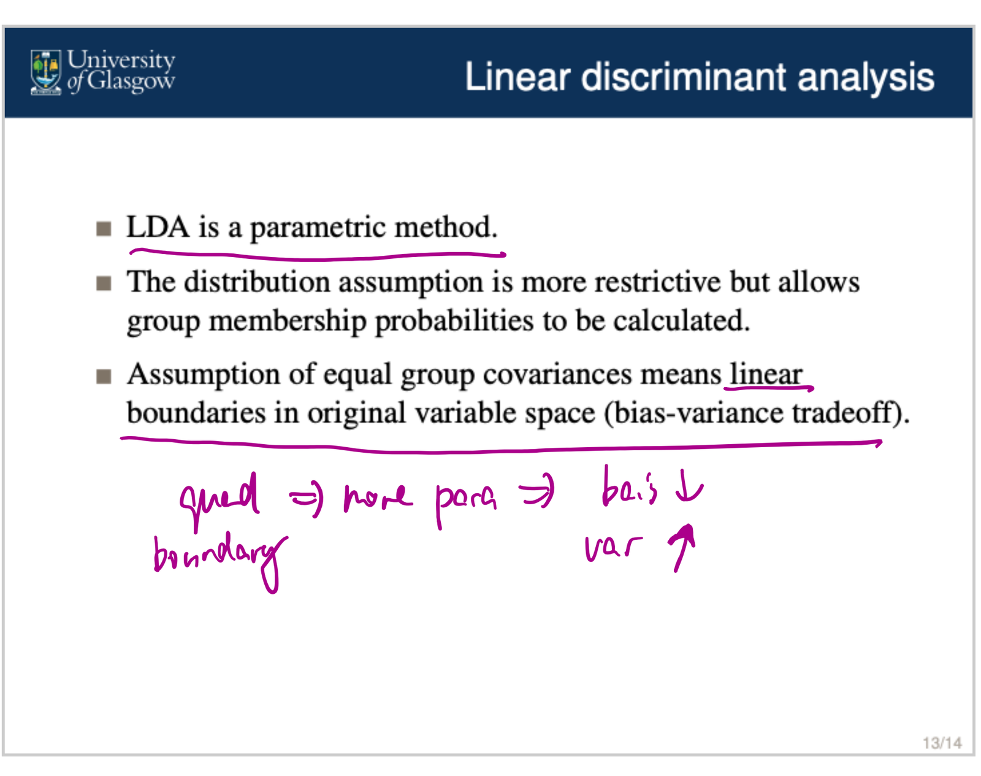


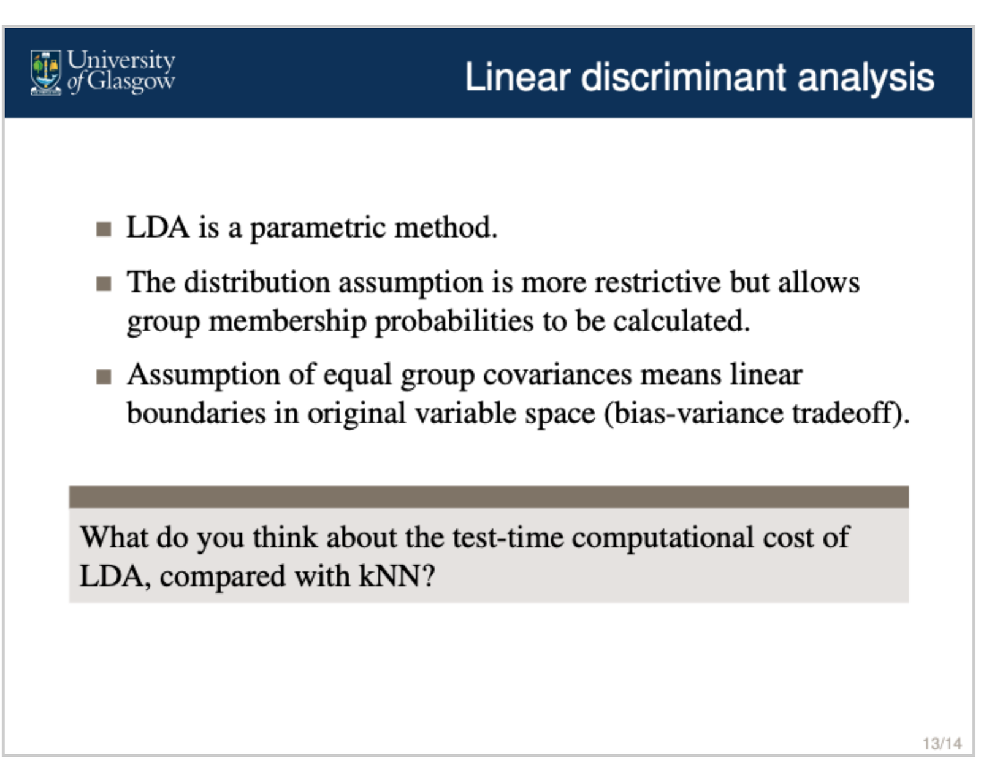


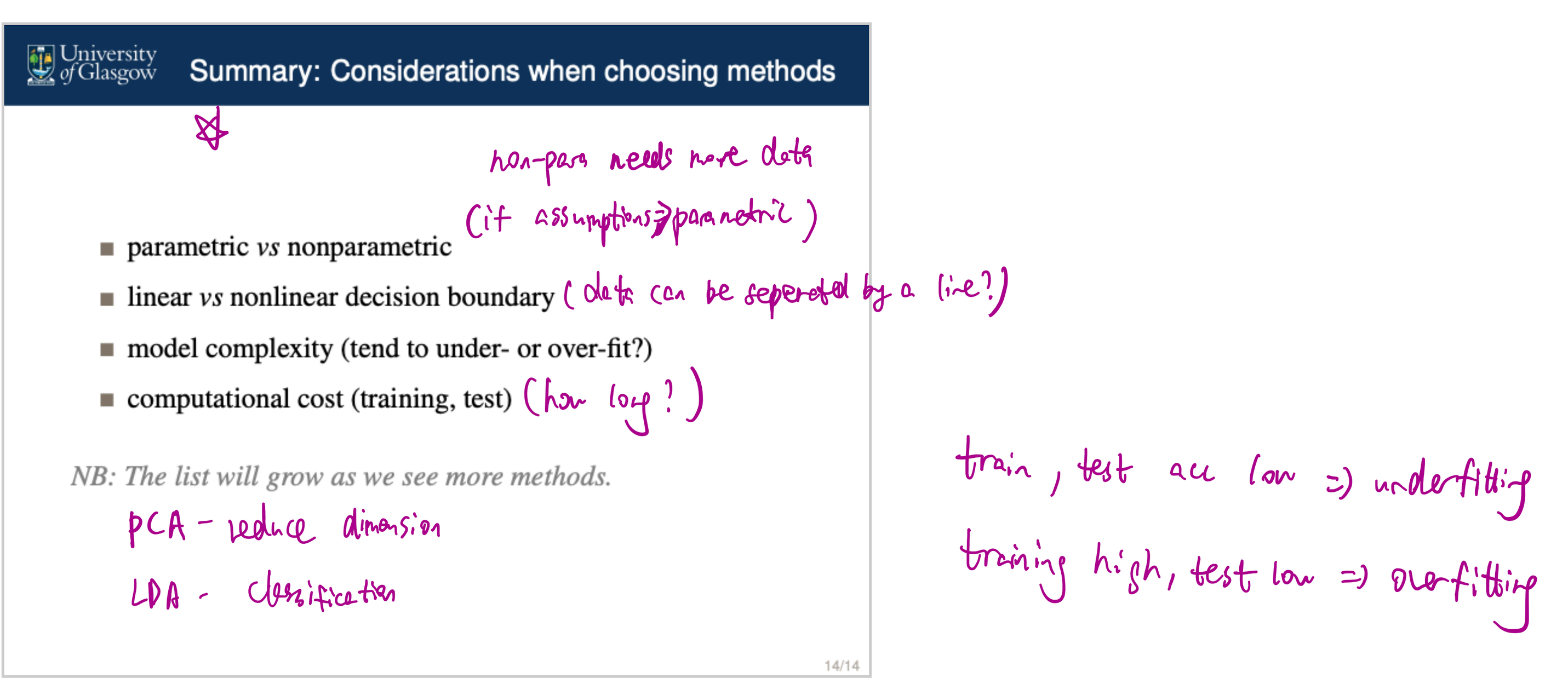

kNN: conceptual question
Suppose that we take a dataset, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use linear discriminant analysis and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest 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
training 0% always itself
test 36% always be the label
> 30
prefer LDA
10/14


kNN: summary
Pros:
Nonparametric (no assumptions on model/data)
Cons:
High test-time computational cost
Break down in high dimensional space¹ (if high dim, do not use knn)
Important considerations:
Value of k (balance between under- and over-fitting) k=1, overfitting, wiggy
k=n, linear boundary
Choice of distances (based on data)
Cont. - Euclidean
category - jacard
NB: The list is not exhaustive.
¹Elements of Statistical Learning, §2.5
11/14


Linear discriminant analysis (LDA)
1 Start with $\Pr(X=\boldsymbol{x}|G=g)$ and assume:
Gaussian distribution: for $\boldsymbol{x}$ in group g, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_g, \Sigma_g)$
Equal covariance: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_G = \Sigma$ (linear decision boundary)
2 Apply the Bayes' theorem
if high ⇒ x ∈ g
$\Pr(G=g|X=\boldsymbol{x}) = \frac{\Pr(X=\boldsymbol{x}|G=g)\Pr(G=g)}{\sum_{r=1}^{G}\Pr(X=\boldsymbol{x}|G=r)\Pr(G=r)}$
3 Bayes discriminant rule: assign $\boldsymbol{x}$ to the class with the largest linear discriminant function
$LDF_g(\boldsymbol{x}) = \boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}_g - \frac{1}{2}\boldsymbol{\mu}_g^\top\Sigma^{-1}\boldsymbol{\mu}_g + \log\pi_g$
12/14


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 boundary ⇒ more para ⇒ bias ↓ var ↑
13/14


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


Summary: Considerations when choosing methods
non-para needs more data
(if assumptions ⇒ parametric)
parametric vs nonparametric
linear vs nonlinear decision boundary (data can be seperated 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 ⇒ underfitting
training high, test low ⇒ 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?