

Department of Statistics

Stats 784 Data Mining

Mid-term test 2017: Friday 22nd September

Answer all 10 questions in the boxes provided. Keep your answers short and succinct. Each question is worth 5 marks.

1. What is the difference between supervised and unsupervised learning?

In supervised learning, we have a data set with a target and several features, and we want to predict the target from the features.
In unsupervised learning, all variables are treated equally and we want to understand the structure of the data cloud.

2. What is the difference between conditional and unconditional prediction error?

Conditional PE is conditional on the training set Z i.e. $PE(x) = E(Y - f_Z(x))^2$
where expectation is taken over Y, X but not Z .
Unconditional also averages over Z i.e.
 $PE = E(PE(x))$

3. What is leave-one-out cross-validation? How does it compare to other type of cross-validation?

leave one out uses n training sets of size $n-1$ and test sets of size 1 i.e. each fold leaves one observation out for a training set.
Has less bias but more variance than the 5 or 10-fold.

4. Why is the 0.632 estimate a better estimate of prediction error than the training error?

Because it is a weighted average of a bootstrapped PE and the training error, so scales the training error up.

5. How do we adjust the complexity of a neural network?

Change the number of hidden layer units.

6. What does the parameter $mtry$ do in a random forest? Why is it important?

For each fitted tree, and each split in that tree, we choose splitting variable from a random subset of size $mtry$ from the set of all variables. This makes the trees less correlated so we get more variance reduction from the averaging.

7. What is the difference between LDA and QDA?

In QDA, we assume the populations have multivariate normal distributions with different covariance matrices. In LDA the covariance matrices are assumed to be the same. For QDA, the boundaries are quadratic, in LDA linear.

8. Describe two loss functions used in classification.

1) $L(y, \hat{y}) = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$

2) $L(y, \hat{\pi}) = \sum_i y_i \log \pi_i$
where $y_i = 1$ if y is class i , and 0 otherwise.

9. Why is it important to standardise features when fitting neural networks and in ridge regression but not in linear regression or logistic regression?

For neural networks, makes starting values work better and the regularization applies to all variables equally. For ridge, so that the regularization applies equally.

In linear regression, scales change the regression coefficients but not the predictor.
Does Logistic.

10. Describe the main features of multiple imputation.

- 1) Fit models to each variable in turn,
- 2) Draw values from the predictive distribution to fill in missing values
- 3) Repeat process to generate several imputed data sets
- 4) Compute predictions for each data set and average