CO902 viva questions

14:00-18:00 10 March 2014

Prepared question

Prepare about 5 minutes of material on one of the following topics:

1. Consider the linear regression model $Y = X\beta + N(0, I_n)$ where X is an $n \times k$ matrix. You can assume that the columns of X are a subset of the columns of an $n \times n$ rotation matrix R. Let $\hat{\beta} = (X^T X)^{-1} X^T Y$. Notation:

   - Total sum of squares $TSS = \sum_{i=1}^{n} Y_i^2$
   - Explained sum of squares $ESS = \sum_{i=1}^{n} (X\hat{\beta})_i^2$
   - Residual sum of squares as $RSS = \sum_{i=1}^{n} (Y_i - (X\hat{\beta})_i)^2$.

   (a) Show that $RSS = TSS - ESS$.

   (b) Show that RSS and ESS are independent. Explain the distributions of RSS and TSS under the assumption that $\beta = 0$.

   (c) Explain the distribution of $(ESS/k)/(RSS/(n-k))$ and how it can be used to test the hypothesis $\beta = 0$.

2. Pick one of the examples of MCMC from `http://www.openbugs.net/w/Examples`. Briefly explain the experimental setting and the probabilistic model for the prior beliefs. Derive a formula for the posterior distribution and explain how you could estimate the mean of the posterior distribution using MCMC. (OpenBugs is a piece of software for doing MCMC, but your answer should not refer to the software specifically.)

3. Describe a publicly available dataset (i.e. one from an R package (but not the `faithful` dataset!!!)) where it would make sense to use k-means clustering. Interpret your results and explain how you choose k.

4. Part of Lab sheet 2, Q3. (The difficulties/dangers of interpreting p-values)

Follow up questions:

1. p-values and confidence intervals.
   (a) Definition of p-values (as random variables) and confidence intervals.
   (b) How to use p-values to form critical regions.
   (c) Constructing a p-value given a test statistic Z with cumulative distribution function F for test of the form $H_0 : \theta = \theta_0$, $H_0 : \theta < \theta_0$, and $H_0 : \theta > \theta_0$.
2. Bayesian statistics:
   (a) The meaning of the terms in the expression "Posterior $\propto$ Prior $\times$ Likelihood".
   (b) The need for Monte Carlo methods due to difficulties in doing exact calculations using the formula above.
3. Importance sampling
4. Rejection sampling
5. Metropolis-Hastings MCMC algorithm. Definition. Implementation issues: Need for burn-in. The use of multiple starting points to check for multi-modal posterior distributions. Visual inspection of the output to assess mixing.
6. SVD: details of the decomposition $X = U\Sigma V^T$. PCA: definition of the principal components in terms of the SVD decomposition of the design matrix X. Interpretation of
   (a) $U\Sigma$ as the PCA transform of X,
   (b) V as a rotation matrix, and
   (c) diagonal$(\Sigma/\sqrt{n})$ as the standard deviations of the principal components.
7. K-means clustering.
   (a) Objective of the algorithm. Given data points $X_1, \ldots X_n \in \mathbb{R}^d$, find k cluster centers $c_1, \ldots, c_k$ such that the cost function below is minimized.
   $$F(c_1, \ldots, c_k) = \sum_{i=1}^{n} \min_{j=1,,\ldots,k} d(X_i, c_j).$$
   (b) Proof that the algorithm approaches monotonically a local minima.