**M.PHIL. IN STATISTICAL SCIENCE**

Friday 30 May 2008 1.30 to 3.30

# INTRODUCTION TO DATA MINING

*Attempt no more than* **THREE** *questions.*

*There are* **FOUR** *questions in total.*

*The questions carry equal weight.*

***STATIONERY REQUIREMENTS***
*Cover sheet*
*Treasury tag*
*Script paper*

***SPECIAL REQUIREMENTS***
*None*

**You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.**

**1**

Consider fitting the additive model $Y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$, $i = 1, \ldots, n$, using the backfitting algorithm.

a. Specify the backfitting algorithm, describing each step. List any identifiability requirements. Assume a generic smoother $S(x)$.

b. Under what circumstances will the answer not be unique? Explain this both mathematically and conceptually.

c. Suppose the algorithm uses a weighted nearest neighbour smoother that puts weight $w$ on the nearest observation, for $0 < w < 1$, and weights $(1 - w)/2$ on the second and third nearest observations. Explain how the bias-variance tradeoff works as a function of $w$.

**2**

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and the ridge regression estimator, $\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

a. Write the objective function for ridge regression and interpret it in the context of shrinkage and multicollinearity.

b. Suppose the prior on $\boldsymbol{\beta}$ is $N(\boldsymbol{0}, \tau^2\boldsymbol{I})$. Show that the posterior mode is the ridge regression estimate, and relate the posterior parameters to $\lambda$.

c. Contrast ridge regression with the LASSO. What are the important differences? Explain the practical implications.

**3**

Consider the use of a weak classifier $G_1(\boldsymbol{x})$, which you may assume may be applied with weights on the observations.

a. Specify the steps in the AdaBoost algorithm.

b. Assume an exponential loss function $L[y, f(\boldsymbol{x})] = \exp[-yf(\boldsymbol{x})]$. Derive the weights in the AdaBoost algorithm.

c. Describe the Random Forest algorithm, and contrast its strategy with AdaBoost.

**4**

Consider the problem of describing model complexity.

a. What is the Vapnik-Červonenkis dimension of a class of models $\{f(\boldsymbol{x}, \boldsymbol{\alpha})\}$ for binary classification? Give an example.

b. For squared error loss, define "in-sample error" as

$$\mathrm{Err}_{\mathrm{in}} = n^{-1} \sum_{i=1}^{n} \mathbb{E}_{Y^N} \mathbb{E}_{\boldsymbol{y}} (Y_i^N - \hat{f}(\boldsymbol{x}_i))^2$$

where $Y_i^N$ represents a new observation at $\boldsymbol{x}_i$ and $\boldsymbol{y}$ denotes the response values in the training data. Also, the training error is

$$\mathrm{err} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{f}(\boldsymbol{x}_i))^2.$$

Define the "optimism" as the expected difference between the in-sample error and the expected training error:

$$\mathrm{Opt} = \mathrm{Err}_{\mathrm{in}} - \mathbb{E}_{\boldsymbol{y}} \mathrm{err}.$$

Show that the optimism is equal to $(2/n) \sum \mathrm{cov}(y_i, \hat{y}_i)$, where $\hat{y}_i$ is the estimated response at $\boldsymbol{x}_i$.

c. Explain multiresolution analysis in the context of complexity for wavelet approximations.

## END OF PAPER