

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 SPECIAL REQUIREMENTS

Cover sheet

None

Treasury tag

Script paper

**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 $Y_i = \beta_0 + \sum_{j=1}^p f_j(x_{ij}) + \epsilon_i$, $i = 1, \dots, n$, using the backfitting algorithm.

- Specify the backfitting algorithm, describing each step. List any identifiability requirements. Assume a generic smoother $S(x)$.
- Under what circumstances will the answer not be unique? Explain this both mathematically and conceptually.
- 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 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and the ridge regression estimator, $\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

- Write the objective function for ridge regression and interpret it in the context of shrinkage and multicollinearity.
- Suppose the prior on $\boldsymbol{\beta}$ is $N(\mathbf{0}, \tau^2\mathbf{I})$. Show that the posterior mode is the ridge regression estimate, and relate the posterior parameters to λ .
- Contrast ridge regression with the LASSO. What are the important differences? Explain the practical implications.

3

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

- Specify the steps in the AdaBoost algorithm.
- Assume an exponential loss function $L[y, f(\mathbf{x})] = \exp[-yf(\mathbf{x})]$. Derive the weights in the AdaBoost algorithm.
- 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 $\{f(\mathbf{x}, \boldsymbol{\alpha})\}$ for binary classification? Give an example.
- b. For squared error loss, define “in-sample error” as

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

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

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

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

$$\text{Opt} = \text{Err}_{\text{in}} - \mathbb{E}_{\mathbf{y}} \text{err}.$$

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

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

END OF PAPER