Friday 24th June 2016
9.30 am – 11.30 am
(2 hours)

ADVANCED STATISTICAL INFERENCE

Answer all questions.

This examination paper includes four questions and is worth a total of 80 marks.

1. Regularized least squares is a popular extension to the standard least squares regression method.

(a) Within this context, describe what is meant by *regularization*.

[2]

Solution: Imposing a constraint on the complexity of the resulting function [2].

(b) The regularized loss for a linear model with regularization parameter λ is given by:

$$\mathscr{L} = (\mathbf{t} - \mathbf{X}\mathbf{w})^{\mathsf{T}} (\mathbf{t} - \mathbf{X}\mathbf{w}) + \mathbf{w}^{\mathsf{T}}(\lambda \mathbf{I})\mathbf{w},$$

where $\mathbf{t}$ is an $N \times 1$ vector of training targets, $\mathbf{X}$ is an $N \times D$ data matrix and $\mathbf{w}$ is a $D \times 1$ vector of model parameters. Show that the set of parameters that minimize this loss, $\widehat{\mathbf{w}}$ is given by:

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{t}$$

[7]

Solution:

$$\begin{aligned}
\text{loss [1]} \quad & \mathscr{L} = (\mathbf{t} - \mathbf{X}\mathbf{w})^{\mathsf{T}} (\mathbf{t} - \mathbf{X}\mathbf{w}) + \mathbf{w}^{\mathsf{T}}(\lambda \mathbf{I})\mathbf{w} \\
\text{multiply out [1]} \quad & = \mathbf{t}^{\mathsf{T}}\mathbf{t} - 2\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \mathbf{w}^{\mathsf{T}}\mathbf{x}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathsf{T}}(\lambda \mathbf{I})\mathbf{w} \\
\text{differentiate [1]} \quad & \frac{\delta \mathscr{L}}{\delta \mathbf{w}} = -2\mathbf{X}^{\mathsf{T}}\mathbf{t} + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} + 2\lambda \mathbf{I}\mathbf{w} \\
\text{equate to zero [1]} \quad & 0 = -2\mathbf{X}^{\mathsf{T}}\mathbf{t} + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} + 2\lambda \mathbf{I}\mathbf{w} \\
\text{re-arrange [1]} \quad & \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \lambda \mathbf{I}\mathbf{w} = \mathbf{X}^{\mathsf{T}}\mathbf{t} \\
\text{factories [1]} \quad & (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \mathbf{I})\mathbf{w} = \mathbf{X}^{\mathsf{T}}\mathbf{t} \\
\text{invert [1]} \quad & \mathbf{w} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}
\end{aligned}$$

(c) Describe how you would go about showing that this was a global minimum.

[3]

Solution: Differentiate the loss again to get the matrix $\frac{\delta^2 \mathscr{L}}{\delta \mathbf{w} \delta \mathbf{w}^{\mathsf{T}}}$[1]. If this matrix at the solution $\widehat{\mathbf{w}}$ is positive (semi)-definite, the solution is a minimum [1]. If this matrix is always positive (semi)-definite, it is a global minimum [1].

(d) State a technique that you might use to choose a value for λ.

[2]

Solution: e.g. Cross-validation [2].

(e) Assume that an oracle tells you that the optimal value of λ is 2. Describe (with diagrams if you like) the effect of setting:

(i) $\lambda \ll 2$

[2]

(ii) $\lambda = 2$

[2]

(iii) $\lambda \gg 2$

[2]

Solution: 2 marks in each case for showing over-fitting to the data, a sensible trade-off between fitting signal and ignoring noise, and underfitting. Also accept mathematical treatment - i.e. effect on $\widehat{\mathbf{w}}$ of these settings.

2. Consider the K-means algorithm, and assume the the number of clusters K is given.

(a) Report a sketch of the algorithm

[10]

Solution: Repeat until convergence:

1. Randomly initialize $\mu_1, \mu_2, \ldots, \mu_K$ [2]
2. Assign each $\mathbf{x}_n$ to its closest μ_k [2]
3. Set $z_{nk} = 1$ if $\mathbf{x}_n$ is assigned to μ_k (0 otherwise) [2]
4. Update μ_k according to: [4]

$$\mu_k = \frac{\sum_{n=1}^{N} z_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} z_{nk}}$$

(b) Show how kernels can be used within the K-means algorithm.

[6]

Solution: The squared distance between data points and centers is redefined from [3]:

$$||\mathbf{x}_n - \mu_k||^2 = \mathbf{x}_n^\mathsf{T}\mathbf{x}_n - 2N_k^{-1}\sum_{m=1}^{N} z_{mk}\mathbf{x}_m^\mathsf{T}\mathbf{x}_n + N_k^{-2}\sum_{m,l} z_{mk}z_{lk}\mathbf{x}_m^\mathsf{T}\mathbf{x}_l$$

to [3]:

$$k(\mathbf{x}_n, \mathbf{x}_n) - 2N_k^{-1} \sum_{m=1}^{N} z_{mk} k(\mathbf{x}_n, \mathbf{x}_m) + N_k^{-2} \sum_{m,l=1}^{N} z_{mk} z_{lk} k(\mathbf{x}_m, \mathbf{x}_l)$$

(c) Discuss when kernels in K-means can be useful from a modeling perspective [2] and discuss at what cost this comes [2].

[4]

Solution: The K-means algorithm cannot deal with nonlinear clusters and the introduction of kernels has the potential to overcome this limitation [2]. Kernels allow to deal with nonlinearity at the cost of introducing extra parameters [2].

3. Classification algorithms can be split into probabilistic and non-probabilistic approaches.

(a) Describe the principle difference (in terms of output) between probabilistic and non-probabilistic classifiers.

[2]

Solution: Probabilistic classifiers provide probabilities of class memberships in their predictions. Non-probabilistic classifiers provide hard classifications.[2]

(b) Give an example of a probabilistic and non-probabilistic classifier.

[2]

Solution: Non-probabilistic [1 for any]: SVM, decision tree, K-nearest Neighbors. Probabilistic [1 for any]: Gaussian process, logistic regression, naive Bayes.

(c) Describe two characteristics of a classification problem that would make raw accuracy (proportion of test points classified correctly) a bad measure of performance.

[4]

Solution: Class imbalance – many more objects of one class than another [2]. Uneven cost of errors (e.g. in clinical applications) [2].

(d) For one of the characteristics given in part (c), describe an alternative evaluation strategy that would overcome the limitation of raw accuracy.

[4]

Solution: Depends on answer to (c) but e.g. Computing the ROC curve and calculating the AUC. Or quoting sensitivity and specificity.

(e) Avoiding over-fitting is a key step in successful classifier design. Describe a procedure that you could follow to determine whether or not over-fitting was a problem in your classifier.

[4]

Solution: Subdivide the data (cross-validation or just extract a validation set) [1]. Train the classifier and optimize parameters on one of the data sets [1]. Test classifier on the other data [1]. If test performance is much worse than training, it is likely that the classifier is over-fitting [1].

(f) In datasets with large numbers of variables, feature selection is often used to overcome the problem of over-fitting. Describe a possible feature selection procedure you might use when faced with such data. Include discussion of how you might determine the number of features to use.

[4]

Solution: Marks [2] awarded for either a procedure by scoring features and then filtering (e.g. z-score) or projection / combination (e.g. PCA).
Marks [2] awarded for describing sensible method for choosing number. E.g. for filtering, cross-validation and picking the number with the best test performance or e.g. PCA, choosing the set of features that accounts for say 95% of the variability in the original data.

4. Consider a classification task based on the Bayesian logistic regression model with parameters $\mathbf{w}$, labels t_n and input data $\mathbf{x}_n$ with $n = 1, \ldots, N$.

(a) Write the likelihood of data given model parameters.

[8]

Solution: Independence of the labels given the parameters is assumed, so the likelihood factorizes over labels [4]:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w})$$

The individual factors of the likelihood read [4]:

$$P(t_n = 1|\mathbf{x}_n, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathsf{T}}\mathbf{x}_n)}$$

$$P(t_n = 0|\mathbf{x}_n, \mathbf{w}) = 1 - P(t_n = 1|\mathbf{x}, \mathbf{w})$$

(b) Assume that a prior $p(\mathbf{w})$ is assigned to model parameters; write the posterior of model parameters given data using Bayes' theorem.

[4]

Solution: The application of Bayes' theorem yields [4]

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

(c) Discuss why the posterior of model parameters given data is analytically intractable.

[2]

Solution: It is not possible to find a functional form for the prior that is conjugate with the likelihood.

(d) When making predictions with logistic regression models, discuss the effect of integrating out parameters compared to the use of a point estimate of parameters.

[2]

Solution: By integrating out parameters, predictions take into account all possible values of the parameters weighted by their posterior probability. This results in a better quantification of prediction variance compared to using point estimates, where only the values of the parameters that maximize the likelihood would be used [2].

(e) What is the idea behind the Laplace approximation to obtain a tractable formulation of the logistic regression model [2] and how does that work [2] (no equations are required)?

[4]

Solution: The idea is to approximate the posterior distribution of parameters given data by means of a tractable distribution, and in particular by means of a multivariate Gaussian distribution [2].

The approximation is done by placing the mean of the Gaussian at the mode of the posterior distribution and by imposing an inverse covariance equal to the negative hessian of the posterior distribution [2].