THE UNIVERSITY of ADELAIDE

Final Examination, Semester 2, 2013

Introduction to Statistical Machine Learning

Official Reading Time:	10 mins
Writing Time:	120 mins
Total Duration:	130 mins

Answer all **5** questions

Instructions

- Begin each answer on a new page in the answer book.
- Examination material must not be removed from the examination room.

Materials

- Calculator without alphanumeric memory or remote communications capability permitted.
- Handwritten lecture notes and printed slides permitted.

DO NOT COMMENCE WRITING UNTIL INSTRUCTED TO DO SO

Question. 1: **Choose the best answer from multiple choices**

1. Cross-validation is a method to:

(a) Remove the curse of dimensionality

(b) Determine the model's sensitivity to noise

2. Principal Component Analysis (PCA) is a method for:

(a) Classification

(b) Reduction of the dimensionality

(c) Probability estimation

3. Which of the following statements is best practice in ML?

(a) Use all the data available for training to obtain optimal performance.

(b) Use all the data available for testing the performance of your algorithm.

(c) Perform cross-validation on training, validation and testing sets

(d) Perform cross-validation solely on training set.

4. Which of the following statements about Machine Learning is **false**?

(a) Machine learning algorithms often suffer from the curse of dimensionality.

(b) Machine learning algorithms cannot generalise.

(c) Machine learning algorithms are typically sensitive to noise.

Question. 2: (a) Define the **regression** and **classification** learning problems.

(b) Suppose that you would like to build a model to estimate the number y of customers arriving in your store in any given hour, based on certain features $\mathbf{x}$ such as store promotions, recent advertising, weather, day-of-week, etc. If you want to design a machine learning algorithm for such a purpose, will you pose this as a classification problem or regression problem?

Question. 3: **Standard ℓ_1-norm soft margin SVMs**

(a) Write down the standard hard-margin SVMs and the ℓ_1-norm soft margin SVMs (both the primal formulations and Lagrange dual formulations.

(b) Explain why we need soft-margin SVMs. In other words, in what kind of cases, hard-margin SVMs will not work.

(c) Explain why we usually solve the Lagrange dual problems instead of directly working on the SVM primal problems?

Question. 4: **ℓ_2-norm soft margin SVMs**

In the class, we have studied that if our data are not linearly then we need to modify our support vector machine algorithm by introducing a soft margin that must be minimised. Specifically, the formulation we have looked at is known as the ℓ_1 norm soft margin SVM. In this problem, we will consider an alternative method, known as the ℓ_2 norm soft margin SVM. This new algorithm is given by the following optimisation problem (notice that the slack penalties are now squared):

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2}\sum_{i}^{m}\xi_i^2 \tag{1}$$

$$\text{subject to}: \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, i = 1 \cdots, m. \tag{2}$$

(a) Notice that we have dropped the $\boldsymbol{\xi} \geq 0$ constraint in the above ℓ_2 SVM problem. **Show that these non-negativity constraints can be removed.** That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

(b) What is the Lagrangian of the ℓ_2 soft margin SVM optimisation problem?

(c) Minimise the Lagrangian with respect to the primal variables $\mathbf{w}, b, \boldsymbol{\xi}$ by taking the following gradients: $\frac{\partial L}{\partial \mathbf{w}}$, $\frac{\partial L}{\partial b}$, and $\frac{\partial L}{\partial \boldsymbol{\xi}}$, and then setting them equal to 0. Here L is the Lagrangian. Write down these calculations.

(d) What is the Lagrange dual of the ℓ_2 soft margin SVM optimisation problem?

Question. 5: **AdaBoost**

(a) The AdaBoost learning algorithm takes an input dataset $\{(\mathbf{x}_i, y_i) : i = 1, ..., m\}$. Briefly describe the algorithm. What is a weak classifier? What is a strong classifier? How does AdaBoost select weak classifiers?

(b) In AdaBoost, how does one update the weak learner's weight? Write down the formula.

(c) A requirement for the weak learner used in AdaBoost is that at each iteration, the selected best weak learner must perform better than random guess. In other words, the weighted error of the selected weak learner must be less than 0.5. Mathematically explain this.

(d) Design one or more heuristic strategies to make AdaBoost more robust to noise or outliers in training data.