

COMP4670/6467

Introduction to Statistical Machine Learning

Assignment 2

Maximum marks	50
Weight	50% of the total assignment marks
Submission deadline	Wednesday, 30 May 2007, 13:00
Submission mode	On paper to M. Hutter or S. Günter or electronic
Estimated time	2-3 hours per lecture week \approx 20min per mark
Late Penalty	20% per day
Some solutions	30.May'07 13:00-16:00

Qinfeng (Javen) Shi will present solutions to some of the exercises on 30.May'07 13:00-16:00 in the Geology Theatre.

Bioinformatics (by Brian Parker) (11/50)

Read the paper, Ambroise, C and McLachlan, G. J, 2003, "Selection bias in gene extraction on the basis of microarray gene-expression data". PNAS, 99:6262-6266.²

In this exercise we will demonstrate how selection bias can be a source of inaccuracy in microarray studies.

Simulate a microarray with no signal i.e. a matrix of random normal data with 100 observations and 1000 genes. For each of $m = 2, 3, \dots, 30$, choose the top m genes based on a t-statistic (equal-variance version), then classify the data using the logistic regression classifier from SG1, and estimate the error rate using 10-fold CV (based on SG3). Plot the graph of error rate vs m . The true error rate should be 0.5. Did the estimated error rate differ? and if so explain why. How would you design a better cross-validation estimate?

Show your plots and the code used to generate them.

²<http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=124442&blobtype=pdf>

Reinforcement Learning (by Doug Aberdeen) (12/50)

Here is a three state Markov decision process:

$$P(s'|s, a = 1) = \begin{bmatrix} 0.2 & 0.8 & 0.0 \\ 0.8 & 0.0 & 0.2 \\ 0.2 & 0.8 & 0.0 \end{bmatrix} \quad P(s'|s, a = 2) = \begin{bmatrix} 0.8 & 0.2 & 0.0 \\ 0.2 & 0.0 & 0.8 \\ 0.8 & 0.2 & 0.0 \end{bmatrix}$$
$$\mathbf{r} = \begin{bmatrix} 1 \\ -1 \\ 8 \end{bmatrix} \quad \gamma = 0.8$$

Use any method you like to:

1. determine the optimal policy, i.e., the best action for each state;
2. determine the long-term discounted value for each state under the optimal policy;
3. determine the long-term average reward under the optimal policy.

Include a brief (less than half a page) description of how you solved these questions. Include any code you used, or paper calculations you made.

Kernel Methods and Support Vector Machines (by SVN Vishwanathan)

Problem 1 (4/50): The aim of this problem is to show that the log-partition function of the exponential family is a nice well behaved function. In fact, one can show more than what is required in this problem. The log-partition function is the cumulant generating function of the exponential family, that is, taking higher order derivatives yields higher order cumulants.

1. Recall that the exponential family of distributions can be written as

$$p(x|\theta) = p_0(x) \exp(\langle \phi(x), \theta \rangle - g(\theta)). \quad (1)$$

2. Write down the analytic formula for $g(\theta)$.
3. Show that $\partial_\theta g(\theta) = \mathbb{E}_{p(x|\theta)}[\phi(x)]$.
4. Show that $\partial_\theta^2 g(\theta) = \mathbf{Cov}_{p(x|\theta)}[\phi(x)]$.
5. Hence conclude that $g(\theta)$ is a convex function of θ .

Problem 2 (6/50): The aim of this problem is to work out the QP that arises from a modified Support Vector regression problem. Suppose instead of fitting a line (in feature space) which minimizes the squared loss we are interested in finding a line which passes through the origin and majorizes all the points in your training sample. This might be useful for instance if you are sampling not only the boundary of a function but also its interior and might arise in many real life applications.

1. Let \mathcal{H} be a dot product space and $\{(\mathbf{x}_i, y_i)\}$ such that $\mathbf{x}_i \in \mathcal{H}$ and $y_i \in \mathbb{R}^+$ for $i = 1, 2, \dots, n$ be a given sample. Let $f := \langle \mathbf{w}, \mathbf{x} \rangle$ for $\mathbf{w} \in \mathcal{H}$ be a class of hyperplanes in \mathcal{H} . Solve the following optimization problem:

$$\min_{\mathbf{w} \in \mathcal{H}} \frac{1}{2} \|\mathbf{w}\|^2 \text{ such that } f(x_i) \geq y_i \quad \forall i \quad (2)$$

2. Kernelize the above problem by using a mapping $k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$.
3. Introduce slack variables into Equation 3 and solve

$$\min_{\mathbf{w} \in \mathcal{H}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \eta_i \text{ such that } f(x_i) \geq y_i - \eta_i \quad \forall i \quad (3)$$

4. Kernelize the slack variable version of the problem.

Bonus points: (+2)

- What is the interpretation of points \mathbf{x}_i such that $f(\mathbf{x}_i) = y_i$?

- Suppose $y_i \in \mathbb{R}$ and we demand $f(\mathbf{x}_i) \geq |y_i|$, does that change the optimization problem significantly?

Problem 3 (5/50): The aim of this problem is to work out the QP that arises from a modified Support Vector classification problem. Suppose we are given a dataset where each error on each sample costs differently. This might be the case, for instance, if we have pre-processed the data and know that it is important to classify certain examples correctly.

1. Assume we have a linear Support Vector Machine with soft margin loss, i.e.

$$c(\mathbf{x}, y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x})) \quad (4)$$

which is to be trained on some training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$. Moreover assume that we know that some of the observations (\mathbf{x}_i, y_i) are more important than others, specifically that there exist weighting coefficients $C_i > 0$ such that we minimize a modified regularized risk functional

$$\sum_{i=1}^m C_i c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \frac{1}{2} \|w\|^2. \quad (5)$$

- (a) Rewrite (5) such that it becomes a constrained quadratic optimization problem, i.e. with linear constraints and a quadratic objective function.
- (b) Derive the Lagrange function corresponding to the constrained optimization problem.
- (c) Compute the dual optimization problem.
- (d) Compare the result to the standard soft margin Support Vector Machine.

Graphical Models (by Tiberio Caetano)

TC1 (3/50) Let $G = (V, E)$ be a Directed Acyclic Graph (DAG) with node set V and edge set E , where $|V| = n$ is the number of nodes. Let $p(x_1, \dots, x_n)$ be a probability distribution defined by $p(x_1, \dots, x_n) := \prod_{i \in V} f_i(x_i, x_{\pi_i})$, where

- π_i is defined as being the set of nodes j such that $j \mapsto i$ is an arrow in the DAG (i.e. the parents of i in the DAG). x_{π_i} is a vector whose coordinates are the variables in π_i .
- $\int_{x_i} f_i(x_i, x_{\pi_i}) = 1, \forall i, x_{\pi_i}$.
- $f_i(x_i, x_{\pi_i}) > 0, \forall i, x_i, x_{\pi_i}$.

Show that $f_i(x_i, x_{\pi_i}) = p(x_i | x_{\pi_i})$. Interpret the result in words.

TC2 (2/50) Explain the concepts of conditional independence and factorization. Explain how directed, undirected graphical models and factor graphs encode (or not) these semantics.

TC3 (2/50) Let G be a triangulated graphical model. How would you compute the marginal probability distribution $p(x_i, x_j)$ if the nodes i and j do not belong to the same maximal clique?

TC4 (2/50) Let G be an undirected graphical model with 7 nodes and with edges corresponding to the node pairs $(1, 2), (2, 3), (3, 4), (1, 5), (3, 5), (5, 6)$. Denote the potential function for maximal clique C by $\psi_c(x_c)$, where c is the set of indexes for the variables in C . Use the elimination algorithm in order to compute $p(x_2)$ and $p(x_5, x_7)$.

TC5 (3/50) Assume you have two types of digital signals: a text and a gray-scale image. Consider letters of the text and pixels of the image as random variables. Now, assume both signals are corrupted by i.i.d. noise (letters in the text have some probability of being flipped to other letters and pixel values have some probability of being drifted within some range). Assume you have the noise models, i.e. the probability distributions according to which the corruption takes place, for each of the two signals. You have lots of *uncorrupted* texts and images available, and in case the noisy text and image didn't have any noise they would have similar statistics to those of the uncorrupted data. Your task is to describe a framework based on graphical models in order to attenuate the noise in the noisy text and image *while retaining the structure of the signals*. For example, you want to remove the noise in the image while retaining its high-order structure, like edges and texture. How would your models look like in both cases (text and image)? Which types of algorithms would you use to remove the noise in each of the signals? How does the computational complexity of your two models (text and image) compare? Can you perform exact inference in these models? Why?