EEL709: Major Test

May 8, 2013

Maximum Marks: 62

Instructions: All working must be clearly shown, with no missing or assumed steps. Whenever words like 'obtain' or 'compute' occur, you should make explicit your entire process for doing so. Your answers should be self-sufficient, not requiring reference to any other materials.

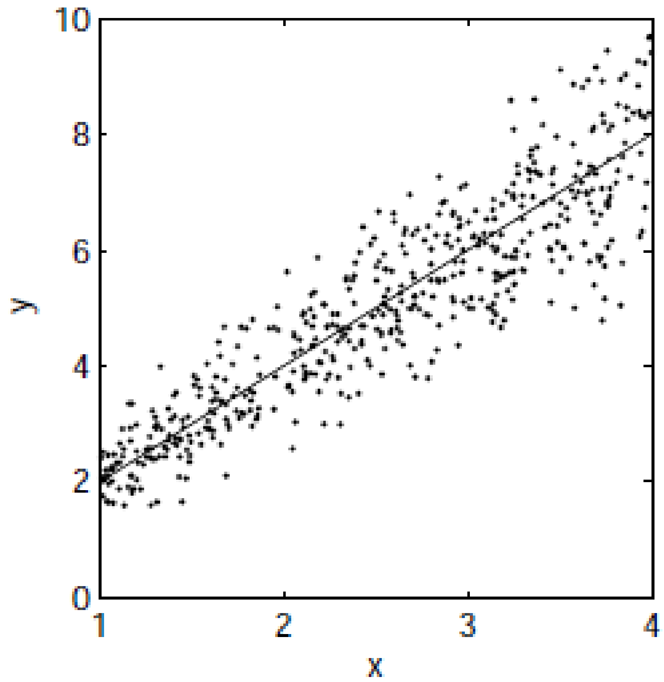
1. (a) Describe, in your own words, the difference between the frequentist and Bayesian views of probability. Can you give an example of a probabilistic statement that makes sense under one interpretation, but not the other? **[3]**

   (b) A coin is tossed 5 times, and 5 heads are observed. How correct would it be to infer that the coin has heads on both sides? Try to answer from both frequentist and Bayesian points of view. **[2]**

2. Here we explore a regression model where the noise variance is a function of the input (variance increases as a function of input). Specifically

$$y = wx + \epsilon,$$

   where the noise ϵ is normally distributed with mean 0 and standard deviation σx. The value of σ is assumed known and the input x is restricted to the interval $[1, 4]$. We can write the model more compactly as $y \sim \mathcal{N}(wx, \sigma^2 x^2)$. If we let x vary within $[1, 4]$ and sample outputs y from this model with some w, the regression plot might look like:



   (a) How is the ratio y/x distributed for a fixed (constant) x? **[2]**

   (b) Suppose we now have N training points and targets $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where each x_n is chosen at random from $[1, 4]$ and the corresponding y_n is subsequently sampled from $y_n \sim \mathcal{N}(wx_n, \sigma^2 x_n^2)$. Obtain the maximum likelihood estimate for w as a function of the training data. **[3]**

(c) Describe in your own words the concept of the bias-variance tradeoff. **[2]**

(d) What are the bias (i.e., difference between expected and actual value) and variance of the estimator for w just obtained, as a function of N and σ^2 for fixed inputs $x_1, ..., x_N$? Can you suggest a method for reducing the variance, even if it involves increasing the bias? **[5]**

(e) Now supposing I put a prior distribution on w: $w \sim \mathcal{N}(0, \alpha^{-1})$, for some fixed α. Obtain the posterior distribution for w, given the same data set as above; also compute the maximum a posteriori estimate. **[3]**

(f) What are the bias and variance of this estimator? What do you infer from this about the role of α in controlling the bias-variance tradeoff? **[5]**

(Some potentially useful relations: if $z \sim \mathcal{N}(\mu, \sigma^2)$, then $az \sim \mathcal{N}(a\mu, \sigma^2 a^2)$ for fixed a. If $z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and they are independent, then $Var(z_1 + z_2) = \sigma_1^2 + \sigma_2^2$.)

3. Here we will look at methods for selecting input features for a logistic regression model

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2).$$

The available training examples are very simple, involving only binary valued inputs:

Number of copies	x_1	x_2	y
10	1	1	1
10	0	1	0
10	1	0	0
10	0	0	1

So, for example, there are 10 copies of $\mathbf{x} = (1, 1)^{\mathrm{T}}$ in the training set, all labeled $y = 1$. The correct label is actually a deterministic function of the two features: $y = 1$ if $x_1 = x_2$ and zero otherwise. We define greedy selection in this context as follows: we start with no features (train only with w_0) and successively try to add new features provided that each addition strictly improves the training log-likelihood. We use no other stopping criterion.

(a) Could greedy selection add either x_1 or x_2 in this case? **[2]**

(b) What is the classification error on the training examples that we could achieve by including both x_1 and x_2 in the logistic regression model? **[2]**

(c) Suppose we define another possible feature to include, a function of x_1 and x_2. Which of the following features, if any, would permit us to correctly classify all the training examples when used in combination with x_1 and x_2 in the logistic regression model: $x_1 - x_2$, $x_1 x_2$, x_2^2? **[3]**

(d) Could the greedy selection method choose this feature as the first feature to add when the available features are x_1, x_2 and your choice of the new feature? **[2]**

4. Given an unlabeled set of examples $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, the *one-class SVM* algorithm tries to find a direction $\mathbf{w}$ that maximally separates the data from the origin. More precisely, it solves the (primal) optimization problem:

$$min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$
$$s.t. \ \mathbf{w}^{\mathrm{T}}\mathbf{x}_n \geq 1 \ \ \forall n = 1, ..., N$$

A new test example $\mathbf{x}$ is labeled 1 if $\mathbf{w}^{\mathrm{T}}\mathbf{x} \geq 1$, and 0 otherwise.

(a) Write down the corresponding dual optimization problem for the above. Simplify your answer as much as possible. In particular, $\mathbf{w}$ should not appear in your answer. **[5]**

(b) Can the one-class SVM be kernelised, both in training and in testing? Why or why not? **[2]**

(c) Can you suggest an application where the one-class SVM might be useful? **[1]**

5. Consider a simple example (due to Judea Pearl), where a burglar alarm at my house (A) can be set off either by a burglary (B) or an earthquake (E). I have two neighbours, John (J) and Mary (M), either of whom could call me in case the alarm goes off.

(a) Draw a Bayesian network to represent the causal relationships between these five binary random variables. **[2]**

(b) Write down the factorisation of the full joint distribution represented by your network. Also specify at least three of the conditional independencies implied by this factorisation. **[2]**

(c) Give an instance in this network of the *explaining away* property, i.e., when a particular variable is observed then another pair of variables which were previously independent, become conditionally dependent. **[1]**

(d) Show that if our model is such that the alarm always (deterministically) goes off whenever there is an earthquake:

$$P(A=1|B=1,E=1) = P(A=1|B=0,E=1) = 1,$$

then $P(B=1|A=1,E=1) = P(B=1)$, i.e., observing an earthquake provides a full explanation for the alarm. **[3]**

6. Consider a setting where, over 5 successive days, when I get back home in the evening I observe the grass on my lawn to be either *wet* or *dry*. Because I work far away from home, I could not observe what the daytime weather was like on those 5 days, but I know that each day it was either *sunny* or *rainy*. Suppose also that I know the following: if the weather was rainy, the probability of the grass being wet in the evening is 0.9; if it was sunny, this probability is 0.3 (there is a sprinkler which the gardener switches on sometimes); if it is rainy one day, then the probability of rain the next day is 0.4; if it is sunny, then the probability of rain the next day is 0.2; and finally, the probability of it being rainy to start with is 0.2.

(a) Draw an appropriate Hidden Markov Model to represent this situation. Specify clearly your notation for random variables, and the corresponding initialisation, emission, and transition probabilities. **[2]**

(b) Suppose my actual observations over the 5 days are $\{wet, dry, dry, wet, wet\}$. Based on this and my specified model, I wish to estimate the probability that the weather on the 5th day was sunny. Use the forward-backward algorithm to compute this. Referring to the notation used in class, which α value(s) do you need to evaluate for this purpose? Show the steps of the recursion involved in doing so. **[5]**

7. Give a concise description of any one of the presented term projects, other than those that were presented on the same day as yours. In addition to outlining the basic methodology and key outcomes, you should briefly mention what you think were the strengths and weaknesses of the approach taken, and what alternative or additional directions for study you would suggest.

(Note: No marks will be given for merely copying stuff from the submitted abstracts.) **[5]**