10-715 Advanced Introduction to Machine Learning

Mid Term Exam *Oct 27th 2014*

Name: : _____

Andrew Id: : _____

Department: : _____

Guidelines

1. **Please don't turn this page until instructed.**
2. Write your name, Andrew Id and department in the space provided above.
3. You have seventy **(70)** minutes for this exam.
4. This exam has **eleven (11)** pages on eleven sheets of paper including this page.
5. This exam has **four (4)** questions. The number of points allocated for each question is indicated next to the question. The total number of points is 80.
6. This exam is **close book**. You cannot use any books, class notes or cheat sheets during the exam.
7. The questions vary in difficulty. The points allocated per question **do not** entirely reflect the level of difficulty. Don't spend too much time on one question.
8. If any question is unclear, you may write your own interpretation and answer the question.
9. The questions appear only on side of the paper. You may use the other side for rough work. If you still need extra paper, please ask one of the instructors.

1 Regression

1.1 Linear Regression and Neural Networks

Your are given some data $(x_i, y_i)_{i=1}^n$, where $x_i = [x_i^{(1)}, \ldots, x_i^{(D)}]^\top \in \mathbb{R}^D$, $y_i \in \mathbb{R}$. There is additive Gaussian noise in the label, i.e. $y_i = f(x_i) + \epsilon$ for some unknown function f and $\epsilon \sim \mathcal{N}(0, \eta^2)$. You are asked to use techniques from linear regression to learn a model to predict y from x. You may assume that $n \gg D$ and that η^2 is small compared to the variation in the y_i values.

You split your training data into two halves: a training set on which you wish to apply your learning algorithm and a test set on which you wish to evaluate your model to choose model hyper parameters.

You begin with a simple linear basis $\phi(x) = [1, x^{(1)}, \ldots, x^{(D)}]^\top$. To estimate w, you minimize a cost function of the form

$$J(w) = \sum_{i=1}^{n} (y_i - w^\top \phi(x_i))^2 \tag{1}$$

For questions 1-5 circle the correct answer (5 $\times$ 2 Points).

1. You notice that both the training error and the test set error are unacceptably high. This is most likely because our model has,
   (a) High bias
   (b) High variance
2. You realize that a linear basis is not rich enough for the problem. So you use a quadratic basis containing all first and second order interactions as shown below,

   $$\phi(x) = [1, x^{(1)}, \ldots, x^{(D)}, {x^{(1)}}^2, \ldots, {x^{(D)}}^2, x^{(1)}x^{(2)}, x^{(1)}x^{(3)}, \ldots, x^{(D-1)}x^{(D)}]^\top \tag{2}$$

   When compared to the linear basis, when using the quadratic basis
   (a) The bias increases but the variance decreases.
   (b) The bias decreases but the variance increases.
   (c) Both the bias and variance increase.
   (d) Both the bias and variance decrase.
3. You realize that the quadratic basis achieves good training error but poor test error. So you modify the cost function with a quadratic penalty: $J(w) = \sum_{i=1}^n (y_i - w^\top \phi(x_i))^2 + \lambda \|w\|^2$. Now, the training error increases slightly but the test set error decreases significantly. This is because by introducing a penalty,
   (a) We reduce both the variance and the bias.
   (b) We increase the variance slightly but significantly reduce the bias.
   (c) We increase the bias slightly but significantly reduce the variance.
4. Consider the prediction y_* at a point x_*.
   (a) y_* is a linear combination of the labels $(y_i)_{i=1}^n$ when we use a linear basis but this is not the case when we use a quadratic basis.
   (b) y_* is a linear combination of the labels $(y_i)_{i=1}^n$ when we use both the linear and quadratic bases.
   (c) y_* has a complicated nonlinear connection to $(y_i)_{i=1}^n$ as we need to invert a matrix to solve linear regression.

5. Now, instead of using classical regression techniques, you decide to use neural networks to train the model. The multi layer perceptron (MLP) is a commonly used neural network and is trained using the back propagaion (BP) algorihtm. BP performs,
   (a) Gradient descent on the MLP cost function.
   (b) Newton's method on the MLP cost function.

1.2 Multi task Regression

Now you are given K batches of data $(y_{ki}, x_{ki})_{i=1}^{n_k}$, $k = 1, \ldots, K$, $x_{ki} \in \mathbb{R}^D, y_{ki} \in \mathbb{R}$ for all i, k. Therefore you have K tasks. For each of the k tasks, you should fit *different* linear regression models for the data. However, you wish to combine them through a complexity penalty for the parameters. Precisely we wish to minimize an objective of the form,

$$\sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ki} - \theta_k^\top x_{ki})^2 \quad + \quad R(\Theta)$$

where, $\theta_k \in \mathbb{R}^D$ are the parameters for the k^{th} task, $\Theta = [\theta_1^\top; \theta_2^\top; \ldots; \theta_K^\top] \in \mathbb{R}^{K \times D}$ and R is a complexity penalty for Θ.

Below we describe three different situations and then four options for R. To each situation, match an appropriate penalty function. Each penalty function should be matched to **only one** situation. Exactly, one penalty function will not be matched.
You do not need to provide justification.

Situations

1. You know that most of the D features are irrelevant and that the y values for each task can be predicted by a *small unknown* subset of the features. Further, these features are the *same* for all tasks.
2. You know that most of the D features are irrelevant and that the y values for each task can be predicted by a *small unknown* subset of the features. However, these features could be *different* for each task.
3. You know that all D features are relevant for all tasks. However, the amount of data n_k we have for each task is either smaller than or not significantly greater than D.

Penalty Functions

(a) $\sum_{k=1}^{K} \sum_{d=1}^{D} |\Theta_{kd}|$

(b) $\sum_{k=1}^{K} \sum_{d=1}^{D} \Theta_{kd}^2$

(c) $\sum_{k=1}^{K} \sqrt{\sum_{d=1}^{D} \Theta_{kd}^2}$

(d) $\sum_{d=1}^{D} \sqrt{\sum_{k=1}^{K} \Theta_{kd}^2}$

Give your matchings in the table below. **(3 × 2 Points)**

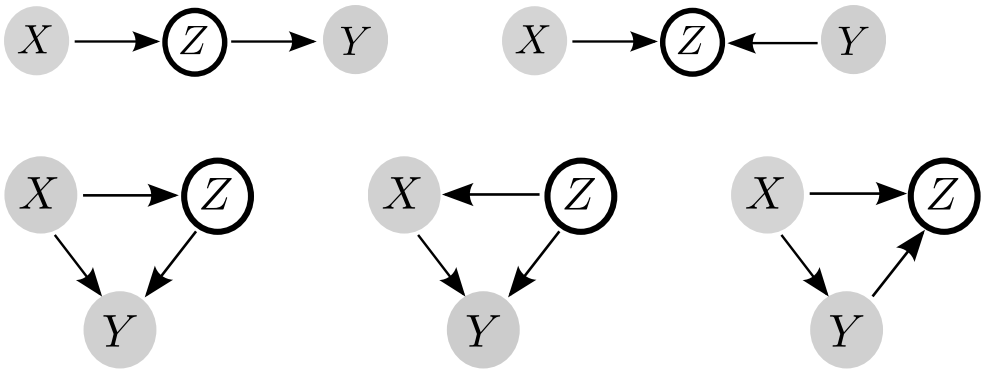
Situation	Penalty Function
1	.....
2	.....
3	.....

2 EM for Supervised Learning

In class and in the homeworks we applied EM for unsupervised learning. Here we will look at EM for a classification problem.

Consider the following problem in astrophysics. We want to assign a "size" label Y to a star indicating if it is a dwarf ($Y = 0$) or giant ($Y = 1$) using its spectrum X observed through telescopes. For this task we have training data $(X_i, Y_i)_{i=1}^n$. However, using X alone we cannot build a good classifier. Astronomers group stars into spectral classes Z such as brown, red, white. Note that the spectral class Z is *not* observed. However, it can be determined using the spectrum X. Further, the size labels Y can be determined using *both* X and Z. We wish to incorporate this domain knowledge when building our classifier.

1. Circle the Bayesian network below that best captures the above scenario. We have shaded the observed variables. **(2 Points)**



Now for simplicity, we will assume that $X \in \mathbb{R}^D$ and both Z and Y are binary valued. We will not make any assumption on the distribution of X. We will use logistic regression to model Y and Z as explained below.

Recall that in logistic regression we model the label V given the features U with parameter β via,

$$p(v|u;\theta) = \sigma(\beta^\top u)^v (1 - \sigma(\beta^\top u))^{1-v}$$

where σ is the logistic function[1].

In this problem, we will model $Z|X$ via logistic regression with parameter θ and $Y|X, Z = i$, (where $i = 0, 1$) via logistic regression with parameter α_i. The parameters of the model are $\theta, \alpha_0, \alpha_1 \in \mathbb{R}^D$.

2. **(3 Points)** First write down an expression for the conditional probability $p(y|x)$. Use this to write down the conditional log likelihood of the parameters $\theta, \alpha_0, \alpha_1$.

Space provided on the next page.

[1] $\sigma(t) = \frac{1}{1+e^{-t}}$. But you can write your answers in terms of σ

cont'd ..

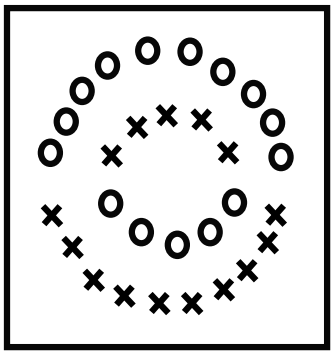
3. **(1 Point)** Are we using a discriminative model or a generative model for classification ?

4. **(3 Points)** We can learn the parameters of the model via Expectation Maximization. Write down the E step update at the $(t+1)^{th}$ iteration. Your answer should be in terms of the estimates $\theta^{(t)}, \alpha_0^{(t)}, \alpha_1^{(t)}$ at the t^{th} iteration. **You do not need to derive the M-step.**

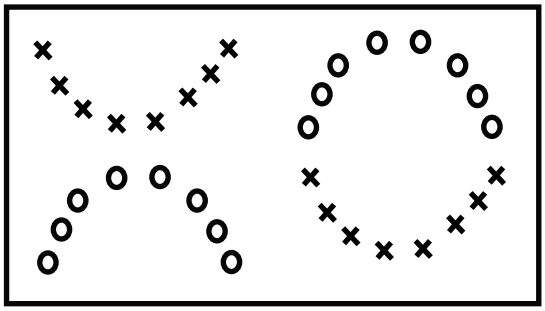
Space provided on the next page.

cont'd ..

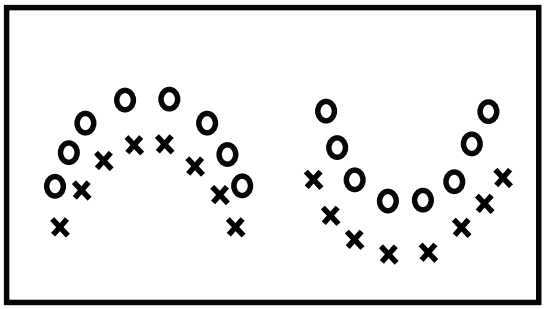
5. **(5 Points)** Consider the following statements about the EM algorithm *in general*. State if they are True (T) or False (F).
   (a) The EM Algorithm maximizes a lower bound on the log likelihood at each iteration.
   (b) The EM Algorithm minimizes an upper bound on the log likelihood at each iteration.
   (c) The EM algorithm *never* decreases the log likelihood.
   (d) EM is guaranteed to converge to a unique globally optimal solution.
   (e) It is *impossible* to use gradient descent/ ascent to learn the parameters instead of EM.
6. **(3 × 2 Points)** Given below are three different datasets on which we wish to apply this model (without any transformation of the variables). Which datasets can we perfectly classify using this model ? We have indicated the two classes using circles (∘) and crosses (×). A simple Yes/ No answer is sufficient.
   (a)



   (b)



(c)



7. **(4 Points)** Consider the following two decision rules to classify a new point x_* using this model.

   - First choose $z_* = \text{argmax}\, p(z|x_*; \theta)$. Then choose $y_* = \text{argmax}\, p(y|z_*, x_*; \alpha_{z_*})$.
   - Choose $y_* = \text{argmax}\, p(y|x_*; \theta, \alpha_0, \alpha_1)$

   Show that these two decision rules are not equal.
   You do not need to provide a proof but should sufficiently explain why this occurs. One option, is to illustrate via a two dimensional example. You can depict the various logistic regression classifiers via straight lines in your illustration.

3 SVMs, Kernels, RKHS

1. **(2+4+4 points)** Consider the primal soft-margin kernel SVM problem with feature map ϕ:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t } \forall i \in [n], y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

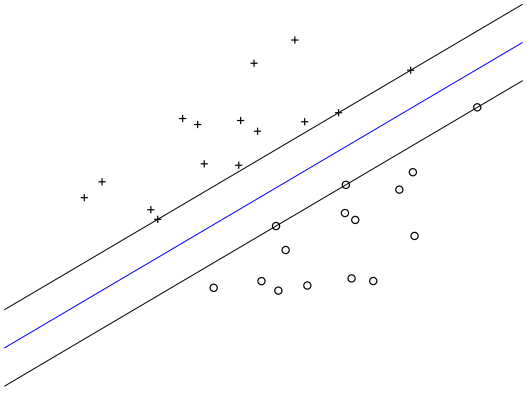
   (a) Why do we usually solve the dual of this problem? Given **one** reason.

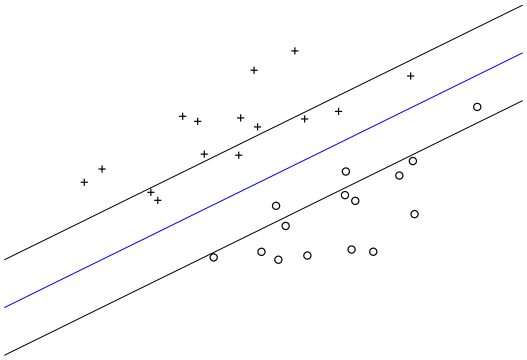
   (b) Suppose $\alpha = [5, 1, 0.99, 0.1, 0, 0, 0, 0, 0.1, 0.99, 1, 5]$ is the solution to the dual, when trained with the data $\{(x_i, y_i)|i = 1, \cdots, 12\}$. The regularization parameter C is not known to us.

      i. List all the support vectors.

      ii. List the training points that are definitely not misclassified using the decision function $f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(x_i), \phi(x)\rangle + b$ formed by α. You can read off this list by simply looking at α.

   (c) The left figure below shows the decision boundary with margin of an SVM trained on some labeled data with regularization parameter C. Circle the data points corresponding to the support vectors. The right figure shows the same with regularization parameter C'. Is C' larger or smaller than C? Justify in one sentence.





2. **(2 points)** Can $g(u, \lambda) = e^{u-\lambda}$ be the Lagrange dual function of some minimization problem? Justify in one/two sentences.

cont'd..

3. **(5 points)** Consider the labeled data set $\{(0,+),(1,-),(2,+)\}$, with the first coordinate denoting the feature and the second coordinate denoting the label. Find a polynomial kernel such that the data is linearly separable under a feature map of the kernel.

4. **(3 points)** What is the dimension of the RKHS associated with the linear kernel $k(x,y) = \langle x, y \rangle$ for $x, y \in \mathbb{R}^d$?

5. **(2 points)** In one sentence, what power does the Representer theorem offer when we are optimizing over infinite dimensional spaces?

4 Graphical Models

1. **(1+2+2 points) Representation**

   Suppose $X_1, X_2, \cdots, X_T$ are T binary random variables in a time series. Assume $T > 2$.

   (a) How many parameters does their joint distribution have, without any assumptions on the relations between them?

   Suppose a domain expert tells us that the value of X_{i+2} is influenced only by the previous two variables X_{i+1}, X_i for $i = 1, 2, \cdots, T-2$ and that the effect of any other previous variables on X_{i+2} can be ignored and further that X_2 is influenced only by X_1.

   (b) Draw a directed graphical model incorporating this assumption.

   (c) How many parameters do we need to specify the joint distribution under this assumption?

2. **(1+2+2+1+2 points) HMMs, CRFs**

   (a) In HMMs, what algorithm would you use to decode a sequence of hidden states $y_1, \cdots, y_T$ given the observed sequence $x_1, \cdots, x_T$?

   (b) Can we use posterior decoding to do this decoding? Justify in a sentence. Recall that in posterior decoding, we infer the most likely hidden state at time t, one t at a time.

(c) What is the Baum-Welch algorithm used for? Is it an EM algorithm?

(d) True/False: HMM is a discriminative model whereas Conditional Random Field(CRF) model is a generative model. Justify in a sentence.

(e) How does the CRF model overcome the label bias problem in Maximum Entropy Markov Model?

3. **(2+2+1 points) Markov Chain Monte Carlo(MCMC)**

(a) Why is detailed balance a desirable property in MCMC sampling?

(b) Why do we need to discard the first several samples in MCMC?

(c) What is the transition kernel in Gibbs sampling?