Machine Learning Midterm

This TWO-SIDED exam is open book. You may bring in your homework, class notes and text- books to help you. You will have 1 hour and 15 minutes. Write all answers in the blue books provided. Please make sure YOUR NAME is on each of your blue books. Square brackets [] denote the points for a question. ANSWER ALL FOUR QUESTIONS FOR FULL CREDIT

1. **Decision Trees**

   Instead of entropy in decision trees, one can use the Gini index or the Miss-classification index.

   (a) [15] You have a data set with 400 positive examples and 400 negative exapmles. Denote this as $(400^+, 400^-)$. Now suppose you have two possible splits for a decision tree. One branch results in $(300^+, 100^-)$ and $(100^+, 300^-)$. The other choice of feature results in $(200^+, 400^-)$ and $(200^+, 0^-)$.

   i. What is the decrease in impurity for the Gini index for each of these choices?

   ii. What is the decrease in impurity for the Missclassification index?

   (b) [10] Which of the two methods of splitting would be preferred and why?

   Gini index: $F(n) = \sum_{l=1}^{k} p_l(n)(1 - p_l(n))$
   Missclassification index: $F(n) = (1 - \max_{l \in [1,k]} p_l(n))$

   The $(300^+, 100^-)$ and $(100^+, 300^-)$ split:
   Missclassification:

$$\widetilde{F}(n) = \frac{1}{2} - \left[\frac{1}{2}\left(\frac{1}{4}\right) + \frac{1}{2}\left(\frac{1}{4}\right)\right] = \frac{1}{4}$$

   Gini:

$$\widetilde{F}(n) = \frac{1}{2} - \left[\frac{1}{2}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + \frac{1}{2}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\right] \times 2 = \frac{1}{8}$$

   The $(200^+, 400^-)$ and $(200^+, 0^-)$ split:
   Missclassification:

$$\widetilde{F}(n) = \frac{1}{2} - \left[\frac{3}{4}\left(\frac{1}{3}\right) + \frac{1}{4} \times 0\right] = \frac{1}{4}$$

   Gini:

$$\widetilde{F}(n) = \frac{1}{2} - \left[\frac{3}{4}\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + \frac{1}{4} \times 0\right] = \frac{1}{6}$$

   Gini prefers the purity split; Missclassification does not care.

2. **Dual Problem**

(a) [5] When given a problem of solving

$$Ax = y$$

where A is not of full rank, one solution is to regularize the system by charging for the length of x, so that:

$$E = ||Ax - y||^2 + \lambda||x||^2$$

Show that this approach implies that

$$A^T Ax + \lambda Ix = A^T y \tag{1}$$

(b) [5] If this equation is rewritten in the form

$$x = \frac{1}{\lambda}(A^T y - A^T Ax) = A^T \alpha \tag{2}$$

what is the equation for α?

(c) [15] Use equations (1) and (2) to eliminate x and create a dual problem involving only A, y, and α and solve it for α.

$$E = [Ax - y]^T[Ax - y] + \lambda x^T x$$

$$E_x = 0$$

$$A^T[Ax - y] + \lambda x = A^T y$$

$$x = \frac{1}{\lambda}[A^T y - A^T Ax] = A^T[\frac{1}{\lambda}[y - Ax] = A^T \alpha$$

Eliminate x,

$$A^T[AA^T + I\lambda]\alpha = A^T y$$

$$\alpha = [AA^T + I\lambda]^{-1} y$$

3. **VC Dimension**

   (a) [5] What is the value of the VC dimension to machine learning? Be VERY BRIEF in your response.

   (b) [10] Consider the function set of axis-parallel rectangles for a two dimensional data set of pairs of real numbers. What is the VC dimension of this set?

   (c) [10] Consider the function set of Support Vector Machine classifiers. What can you say about the VC dimension of SVMs?

(a) It allows you to bound the error on the test set.

(b) By construction $d = 4$.

(c) $d + 1$ where d is the dimension of the mapping function vector.

4. **Sampling** The Box-Muller method of generating random numbers from a Gausian ultilizes random numbers drawn from the uniform distribution $[-1, 1] \times [-1, 1]$ and filtered by rejecting points that do not satisfy $z_1^2 + z_2^2 \leq 1$. Then (y_1, y_2) are computed by:

$$y_1 = z_1 \left(\frac{-2lnr^2}{r^2} \right)^{\frac{1}{2}}$$

$$y_2 = z_2 \left(\frac{-2lnr^2}{r^2} \right)^{\frac{1}{2}}$$

where $r^2 = z_1^2 + z_2^2$. The resultant y_1 and y_2 are independent and each have zero mean and unit variance.

   (a) [5] Where $y = (y_1, y_2)$, write down the covariance matrix $cov(y)$, that is, $E[yy^T]$.

   (b) [10] It would be really helpful if we could sample from a Gaussian with an arbitrary mean μ and covariance Σ. It turns out that this can be done, since we can factor Σ as

$$\Sigma = LL^T$$

   Show that $cov(x) = \Sigma$ and $E(x) = \mu$ where

$$x = Ly + \mu$$

   (c) [10] Suppose that we would like to sample from a distribution that can be expressed as a weighted mixture of Gaussians, that is

$$p(x) = w_1 N(\mu_1, \Sigma_1) + w_2 N(\mu_2, \Sigma_2) + w_3 N(\mu_3, \Sigma_3)$$

   How would you design a program to sample from $p(x)$?

(a)

$$E[y] = 0; E[yy^T] = I$$

(b)

$$\begin{aligned} covx &= E[xx^T] - E[x]E[x^T] \\ &= E[(\mu + Ly)(\mu + Ly)^T] - \mu\mu^T \\ &= LL^T \\ &= \Sigma \end{aligned}$$

(c) Since you know how to sample from an arbitrary Gaussian, pick one with odds $w_1 : w_2 : w_3$ and sample from it.