COL 774 Major Exam

Tuesday May 10, 2016

Notes:

- **Time: 7:55 am to 10:05 am. Maximum Points: 36**
- **Each question carries 4 points.**
- **This exam is closed books/notes.**
- **Justify all your answers.**
- **Some questions may be longer/more difficult than others. Use your time judiciously.**
- **Start each answer on a new page.**

1. We would like to cluster a set of m data points $\{x^{(i)}\}_{i=1}^{m}$ into K clusters. Given a set of cluster assigments $C = (C^{(1)}, \cdots, C^{(m)})$, and a set of cluster representatives $\mu = \{\mu_1, \cdots, \mu_K\}$, let us define the following error function:

$$J(C, \mu) = \Sigma_{i=1}^{m} \Sigma_{l=1}^{K} \mathbb{1}\{\mu_{C^{(i)}} = l\} ||x^{(i)} - \mu_l||^2$$

where $\mathbb{1}$ denotes the indicator function. We would like to find an assignment of variables (C, μ) which minimizes $J(C, \mu)$. Show that applying block coordinate descent over the blocks of variables C and μ results precisely in the k-means algorithm which is characterized by the following two steps:
(a) Given the current centroids (cluster reprsentatives), assign each data point to the cluster with the closest centroid.
(b) Given the current cluster assignments, compute the centroid for each cluster as the mean of the points belonging to it.
You should explicitly map each minimization step in your block coordinate descent to the two steps of the k-means algorithm as described above.

2. Assume you have a biased coin with probability of heads given by the parameter ϕ. Consider m independent tosses of this coin resulting in a sequence with p heads and n tails ($p + n = m$). Note that the observed data D here corresponds to the (single) sequence of heads and tails as described above. Let the prior distribution over ϕ be given by $\phi \sim Beta(1, 2)$ [1]. Calculate the expected value $E[\phi^2]$ under the posterior distribution $P(\phi|D)$ in terms of n and p. You should simplify your expression as much as possible. You can use the fact that for k_1, k_2 positive integers, $\int_0^1 \phi^{k_1}(1-\phi)^{k_2} d\phi = \frac{k_1! k_2!}{(k_1+k_2+1)!}$

3. Consider rolling an r-faced die where the probability of the k^{th} ($1 \leq k \leq r$) face showing up is given by p_k. We will use $\mathbf{p} = (p_1, p_2, \cdots, p_r)$ to denote the parameter vector. Let the die be rolled n times and let n_k denote the number of times k^{th} face shows up. Note that $\sum_{k=1}^{r} p_k = 1$ and $\sum_{k=1}^{r} n_k = n$. Find the maximum-likelihood estimate of the parameters by maximizing the log-likelihood $LL(\mathbf{p})$ of the data from first principles. You are not allowed to use the theory of Langragians for this question.

4. Consider the SVM formulation to handle noisy data as done in class:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \\ &y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i; \quad \forall i \\ &\xi_i \geq 0; \quad \forall i \end{aligned}$$

The symbols are as defined in class.

(a) Plot ξ_i on y-axis as a function of $y^{(i)}(w^T x^{(i)} + b)$ on x-axis.

[1] Recall $\phi \sim Beta(\alpha, \beta)$ means that $P(\phi) \propto \phi^{\alpha-1}(1-\phi)^{\beta-1}$.

(b) Using cues from part (a) above, formulate the above SVM problem as an unconstrainted minimization problem, i.e., an optimization problem without any constraints. Hint: Think of getting rid of ξ_i variables and changing your objective function approrpiately.

5. Implement a neural network to represent the following function over 3 Boolean variables $y = (x_1 \vee x_2) \wedge (x_1 \vee \bar{x_2} \vee \bar{x_3}) \wedge (\bar{x_1} \vee \bar{x_2})$. You should use the minimum number of perceptron units (and at most one hidden layer) to implement this function. Use the step function as the thresholding unit. Clearly specify the weights of your network. Next, note that though we started with Boolean inputs, the neural network constructed can be interpreted as implementing a function over three real valued inputs. Draw the decision surface enforced by your network in the $x_1 - x_2$ subspace restricting $x_3 = 1$. Clearly mark the regions classified as positive ($y = 1$) and negative ($y = 0$) by your network. Justify your answer.

6. Consider applying PCA over a set of points $\{x^{(i)}\}_{i=1}^m$ where each $x^{(i)} \in \mathcal{R}^n$. Assume that the data has been noramlized to have zero mean and unit variance. Further, assume that we are projecting that data onto a single dimension u. Recall that in PCA, the desired direction u is the one that maximizes the projected variance of the data. Derive the expression for the projected variance in terms of the empirical co-variance Σ from first principles. Next, formulate the problem of finding the desired vector u as a constrained optimization problem. Think about what constraints you need to impose on the u vector. Use the theory of Langragians to show that u must be an eigenvector of the empirical co-variance matrix.

7. Generalized Linear Models (GLMs) assume that the target variable y (conditioned on x) is distributed according to a member of the exponential family:
$$p(y;\eta) = b(y)exp(\eta y - a(\eta)) \tag{1}$$
where $\eta = \theta^T x$. Further, the hypothesis function is given as $h_\theta(x) = E[y|x]$. Consider a learning problem where the target variable y conditioned on x is distributed as a Poisson random variable. Show that the above learning problem corresponds the class of GLMs. You should explicitly describe the values of $b(y), \eta$ and $a(\eta)$ in terms of the parameters (and form) of the Poisson distribution. Also, derive the expression for $h_\theta(x)$ in terms of θ and x. Using this, write down the expression for the gradient descent update rule of the θ parameters when learning the parameters of the above model under least squares regression. Recall that Poisson distribution is a discrete distribution (parameterized by mean parameter λ) defined as $P(y) = \frac{e^{-\lambda}\lambda^y}{y!}$ where $y \in I^+ \cup \{0\}$, I^+ being the set of positive integers.

8. Consider an instance space over the real line i.e., $x \in \mathcal{R}$. Consider a hypothesis class H where each hypothesis $h_\theta \in H$ takes the form $h_\theta(x) = \mathbb{1}[sin(\pi(1+\theta x)) \geq 0]$. Here, $\mathbb{1}$ is the indicator function. Now, consider a set of m points given by $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ where $x^{(i)} = \frac{1}{2^i}$. Show that these m points can be shattered by H. Note that this result shows that H is a hypothesis class with $VC(H) = \infty$ since m can be arbitrarity large. Hint: Think of constructing a hypothesis h_θ with parameter $\theta = \sum_{i=0}^m r_i$ such that $r_i * x^{(i)}$ is a postive integer if $y^{(i)} = 1$ and 0 otherwise, $\forall i > 0$. Fix $r_0 = 1$.

9. Consider the following dual formulation of the SVM problem:
$$\max_\alpha \frac{1}{2}\sum_{i=1}^m \alpha_i - \sum_{i,j=1}^m \alpha_i \alpha_i y^{(i)} y^{(j)} ({x^{(i)}}^T x^{(j)})$$
$$\sum_{i=1}^m \alpha_i y^{(i)} = 0;$$
$$0 \leq \alpha_i \leq C, \forall i.$$
where the symbols are as defined in class. We can optimize this problem by using coordinate ascent over two variables at a time while fixing others. Without loss of generality, let us assume that the variables we are optimizing over are α_1 and α_2 (and others are fixed). Then, it is not difficult to see that above optimization problem reduces to one of the following form:
$$\max_{\alpha_1} a\alpha_1^2 + b\alpha_1 + c;$$
$$0 \leq \alpha_1, \alpha_2 \leq C;$$
$$\alpha_1 y^{(1)} \alpha_1 + \alpha_2 y^{(2)} = k;$$
Where $a, b, c, k \in \mathcal{R}$ are some constants and $a < 0$. Explain in detail how you would solve the above optimization problem to obtain the optimal values of α_1 and α_2. Note that you need to think about how to handle the constraints while maximizing the quadtratic expression above. You may find it helpful to think about the shape of a quadratic function. You are not allowed to use the theory of Langrangians for this problem. This is the well known SMO algorithm for solving SVM dual.