Pratice Set - Major Exam

Sunday April 30, 2017

1. Suppose we are given a set of points $\{x^{(1)}, \ldots, x^{(m)}\}$ in $\mathcal{R}^n$. Let us assume that we have pre-processed the data to have zero-mean and unit variance in each coordinate. Consider applying PCA on this data. We showed the projected directions in PCA correspond to the eigenvectors of Σ where Σ is the empirical co-variance matrix. Show that the variance along a projected direction is proportional to the corresponding eigenvalue. You should explicitly derive the form of variance along each dimension.

2. Consider a learning problem with n Boolean attributes. Let the hypothesis class be H. Let $c \in H$ be the target concept, and D be a set of m independent, randomly drawn examples from c. A hypothesis is said to be consistent with D if it has zero prediction error on the examples in D. Let H' denote the subset of hypotheses such that $\forall h \in H'$ generalization error $\epsilon(h) > \gamma$. Give an upper bound on the probability that $\exists h \in H'$ such that h is consistent with D. You can use the fact that $P(|\epsilon(h) - \hat{\epsilon}(h)| > \gamma) \leq 2 * \exp(-2\gamma^2 m)$, where the symbols are as defined in the class.

3. Consider fitting an SVM with $C > 0$ to a dataset that is linearly separable. Recall that C is the parameter controlling the relative cost of margin violators in the SVM formulation. Is the resulting boundary guaranteed to separate the two classes? If yes, prove your claim. If not, give a counter example and argue appropriately. What happens as you increase the value of C?

4. Let X be a Boolean random variable. Let $p(X)$ and $q(X)$ be two probability distributions defined over X. Let $H_p(X)$ and $H_q(X)$ denote the entropy of X under the distributions p and q, respectively. Let $H_{pq}(X)$, the cross entropy of X under the p and q, be defined as $H_{pq}(X) = \sum_{x_i} -p(X = x_i) log(q(X = x_i))$ where x_i varies over the values that X can take. KL-divergence between p and q, denoted by $KL(p||q)$, is defined as $KL(p||q) = H_{pq}(X) - H_p(X)$. Show that $KL(p||q)$ is not a symmetric measure (Hint: Try coming up with example distributions where $KL(p||q) \neq KL(q||p)$). Also, derive the range of values that $KL(p||q)$ can take given arbitrary distributions p and q over X.

5. Recall that generalized linear models assume that the target variable y (conditioned on x) is distributed according to a member of the exponential family: $p(y; \eta) = b(y) exp(\eta y - a(\eta))$, where $\eta = \theta^T x$. Given a training set $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, the log-likelihood is given by $l(\theta) = \Sigma_{i=1}^m log \quad p(y^{(i)}|x^{(i)}; \theta)$. Give a set of conditions on $a(\eta)$ which ensure that the log-likelihood is a concave function of θ (and thus has unique maximum). Simplify your set of conditions as much as possible. You can use the fact that a function $f(\theta)$ is concave if the corresponding Hessian matrix $(\nabla_\theta^2 f(\theta))$ is negative semi-definite.

6. Consider a learning problem where the train and test errors are given by ϵ_r and ϵ_t, respectively. Assume that you know from the domain knowledge (and other prior experience) that there is a minimum desired error level, ϵ_d, that can be achieved in this domain. When can you say that the learning model is underfitting (and not overfitting)? Describe in terms of the values of the quantities ϵ_r, ϵ_t and ϵ_d. Also, draw the corresponding learning curve i.e. plot ϵ_r, ϵ_t as well as ϵ_d with varying number of examples in the above scenario.

7. Consider a learning problem with training data given as $(x^{(i)}, y^{(i)})$ pairs such that $x^{(i)} \in R^2$ and $y^{(i)} \in \{0, 1\}$. Assume that some of the class labels i.e. $y^{(i)}$'s are missing (unlabeled). This is the setting for semi-supervised learning. Give an example distribution (i.e. draw the set of points in R^2 with corresponding labels) such that taking into account labeled as well as unlabeled points would potentially learn a better classification model in comparison with learning a model using labeled points only.

8. 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.

9. 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,
$\int_0^1 \phi^{k_1}(1-\phi)^{k_2} d\phi = \frac{k_1! k_2!}{(k_1+k_2+1)!}$

10. Consider clustering on the dataset in Figure 1, where each dot represents a data point in $\mathcal{R}^2$. Assume that there are two clusters and the initial means are the two bold circles. Sketch the clusters k-means would derive given the starting points as in the figure. Does this represent the true clustering as evident from the figure? Is there any setting of the initial means such that k-means would be able to recover the true set of clusters? Argue.

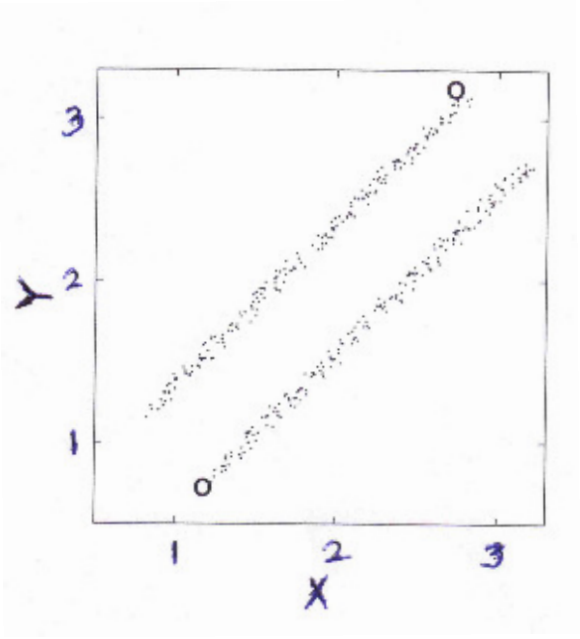

Figure 1: Set of points to be clustered

11. 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.

12. 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 $r_0 = 1$.

13. The EM algorithm that we covered in the class was for solving a maximum likelihood estimation problem. Suppose we are working in a Bayesian framework, and would like to find the MAP estimate of the parameters θ by maximizing

$$\Big(\prod_{i=1}^m p(x^{(i)}|\theta)\Big)p(\theta) = \Big(\prod_{i=1}^m \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}|\theta)\Big)p(\theta)$$

As before, $x^{(i)}$'s denote the observed variables and $z^{(i)}$'s denote the latent variables. $p(\theta)$ is our prior on the parameters. Derive the E(Expectation) and M(Maximization) steps in the EM algorithm to work for the MAP estimation. You may assume that $\log p(x, z|\theta)$ and $\log p(\theta)$ are both concave in θ, and hence, maximizing any linear combination of these quantities is tractable.

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

14. Let the input domain for a learning problem be $X = \mathcal{R}$. Find the VC-dimension of the hypothesis space given by $h(x) = \mathbb{1}\{a * cos(x) > 0\}, a \in \mathcal{R}$,where $\mathbb{1}$ denotes the indicator function. Justify your answer.

15. (**This is only for fun!**) Solve the Picture Puzzle in Figure 2.

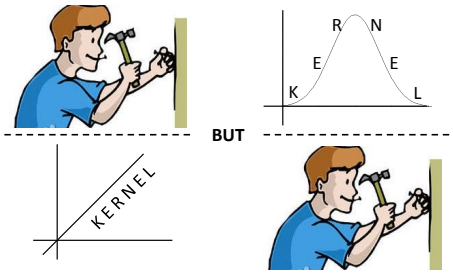

Figure 2: Picture Puzzle