Pratice Set - Major Exam

Monday May 2, 2016

1. Suppose you are minimizing a convex objective function using gradient descent and the algorithm has not converged even after 10,000 steps. What might be the possible reasons? Is there any alternate algorithm (other than gradient descent) that may lead to faster convergence? Argue.

2. 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.

3. Consider a learning problem with n features satisfying the Naïve Bayes assumption i.e. $P(x|y) = \prod_{j=1}^{n} P(x_j|y)$. Let the target variable be $y \in \{0, 1\}$ with $P(y = 1) = \phi$. Let each feature x_j be continuous valued with a Gaussian distribution conditioned on the class variable y. In particular, $P(x_j|y = 0) \sim \mathcal{N}(\mu_{j|0}, \sigma_0^2)$ where $\mu_{j|0}$ is the mean of the distribution and σ_0^2 is its variance. Note that the variance of the distribution (given the class variable y) does not depend on the particular feature x_j. Similarly, we have $P(x_j|y = 1) \sim \mathcal{N}(\mu_{j|1}, \sigma_1^2)$. Above model is called Gaussian Naïve Bayes model. Given the training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, derive the maximum-likelihood estimator for the parameters $\mu_{j|0}$ in the above model. Does your estimate make intuitive sense? Justify. Recall that if $z \sim \mathcal{N}(\mu, \sigma)$, then $P(z) = \frac{1}{\sqrt{2\pi}\sigma} exp^{\frac{(z-\mu)^2}{2\sigma^2}}$.

4. Consider the objective (error) function optimized by linear regression:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))^2$$

   where $h_\theta(x^{(i)}) = \theta^T x^{(i)}$. Show the function $J(\theta)$ is a convex function of the parameters θ, using the property that its Hessian (matrix of second partial derivatives) is positive semi-definite. Note: If you would like to use the matrix form representation for $J(\theta)$, you should explicitly show how you come at this form (and not directly use the result as derived in class).

5. Consider a set of linearly separable data points $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Let the corresponding SVM classifier (separator) be given by the hpyerplane $w^T x + b = 0$. Specifically, given an input x, the classifier outputs 1 if $w^T x + b > 0$ and -1 otheriwse. Now, assume that another training point $(x^{(m+1)}, 1)$ is added to the training dataset such that $(w^T x^{(m+1)} + b) > 1$. Formally prove that addition of this new training point does not change the SVM classifier learned earlier with the original set of m points.

6. Derive the computational complexity of k-means algorithm in terms of m, n, k and T, where m is the number of data points, n is the number of dimensions, k is the number of clusters and T is the number of iterations for which algorithm is run. You should justify the steps of your derivation to get full points.

7. 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. Consider applying PCA on this data. In an alternate formulation of PCA, to find the first principal component, we look for the direction which minimizes the mean squared distance between the projected and the original points. Show that this is equivalent to the original PCA interpretation in which we find the direction maximizing the variance of the projected data. You should not assume any expression for the variance as done in class and derive it from first principles.

8. 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?

9. 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.

10. Consider the Bayesian network shown in Figure 1. The associated CPTs for each node are shown in the figure. Given this network, calculate the posterior probability $P(A = 0|D = 1)$. Note that CPT for $P(E|D)$ is not provided, and hence your computation should not explicitly require the values from this distribution.

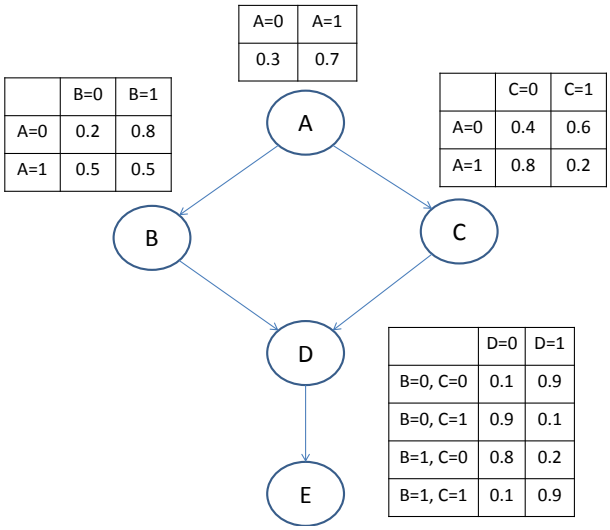

Figure 1: Bayesian Network

11. 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.

12. 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.

13. 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.

14. One way to avoid overfitting in decision trees is to prune the tree using a separate validation set. Typically, a full-blown tree is learnt on the training set first. This is followed by iterative pruning of the learned tree until further pruning does not lead to decrease in error on the validation set. An alternative approach is to keep checking error on the validation set while the tree is being constructed. The tree construction is stopped when the error (on the validation set) does not decrease any further. Which of these approaches do you think would work better in general. Why?

15. Consider the Naïve bayes model with Boolean valued features and binary class labels. Show that $P(y = 1|x)$ takes the form of a logistic function i.e. $P(y = 1|x) = \frac{1}{1+e^{-\theta^T x}}$. Clearly express θ in terms of the parameters of the Naïve bayes model. Do not forget to include the intercept term in θ.

16. Let X and Y be two discrete valued random variables. Let H be the entropy function as defined in class. Let $H(X)$ and $H(Y)$ denote the entropy of the random variables X and Y, respectively. Let $H(Y|X)$ denote the conditional entropy of Y given X. Prove that $H(Y|X) = H(X|Y) + H(Y) - H(X)$. You should prove it from first principles and not use any existing facts about entropy. Note: This is the entropy analagoue of the Bayes rule for probabilities.

17. 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 ($n+p=m$). Note that the observed data D here corresponds to the sequence of heads and tails as described above. Let the prior distribution over ϕ be given by $\phi \sim Beta(2,2)$ [1]. Calculate the expected value $\mathrm{E}[\phi]$ of the parameter ϕ under the posterior distribution $P(\phi|D)$ in terms of n and p. You can use the fact that for k_1, k_2 positive integers,
$\int_0^1 \phi^{k_1}(1-\phi)^{k_2} d\phi = \frac{k_1!k_2!}{(k_1+k_2+1)!}$

18. Let the domain of the inputs for a learning problem be $X = R$. Consider using hypotheses of the following form:

$$h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_d x^d \geq 0\},$$

where $\mathbb{1}$ is the indicator function. Assumed $d \geq 1$. Let $H = \{h_\theta : \theta \in R^{d+1}\}$ be the corresponding hypothesis class. Show that this hypothesis class can not shatter any set of $d+2$ points in the input space. You should prove this from first principles and not use any existing facts about VC dimensions. Hint: Think about how many roots a polynomial of degree d has.

19. Solve the Picture Puzzle in Figure 2. **Note that we may not have any picture puzzles this time around. But this will be fun for you to solve anyway!**

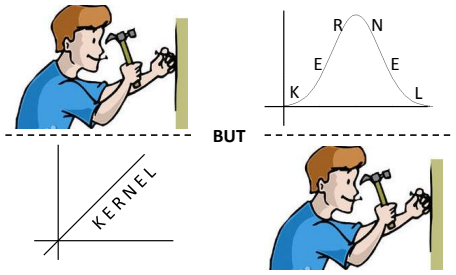

Figure 2: Picture Puzzle

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