COL 774 - Practice Questions

Sunday March 19, 2017

Notes:

1. Given an unlabeled set of examples $\{x^{(1)}, \cdots, x^{(m)}\}$ the *one-class SVM algorithm* tries to maximally separate the data from a hyperplane passing through the origin. More precisely, it solves the following (primal) optimization problem:

$$min_w \frac{1}{2} w^T w$$

$$(w^T x^{(i)}) \geq 1, \forall i = 1, \cdots, m$$

   Derive the dual form for the above optimization problem. Your dual formulation should not contain any primal variables (w). You should simplify your formulation as much as possible.

2. 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?

3. Consider a machine learning problem with input feature vectors $x \in R^n$ and satisfying $||x|| = 1$. Given two vectors $x, z \in \mathcal{R}^n$ (and satisfying above properties), consider the function $K(x, z)$ defined as $K(x, z) = ||x + z||^2$. Show that $K(x, z)$ is a valid Kernel.

4. Draw the decision tree of the smallest height to correctly represent the concept in Figure 1. You are allowed to make only two-way splits over an attribute value i.e. any internal node of the tree will have two children. Further, the only decisions allowed are of the form $X < a$, $X > b$, $Y < c$ and $Y > d$, where $a, b, c, d \in \mathcal{R}$. Make sure to indicate on each branch of the tree whether it corresponds to the condition being *true* or $false$. Also, label each leaf node of the tree appropriately.

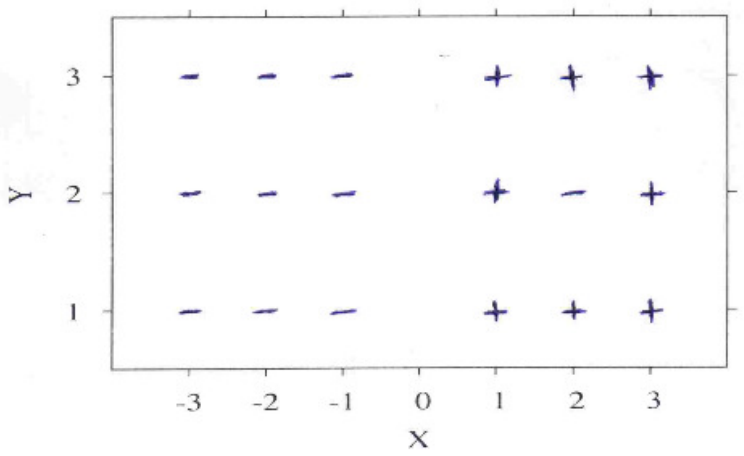

Figure 1: Set of points in two dimensions with corresponding labels.

5. Draw a neural network to represent the Boolean function $f(x_1, x_2, x_3) = (x_1 \vee x_2 \vee x_3) \wedge (\neg x_1 \vee x_2 \vee \neg x_3)$ defined over 3 input variables. Note that $x_1, x_2, x_3 \in \{0, 1\}$. Here, $\wedge$ denotes the *and* operator, $\vee$ denotes the *or* operator and $\neg$ denotes the *negation*, as in the standard Boolean algebra. Your network should have at most one hidden layer and use at most 3 network units. Use the threshold function $g(x) = \mathbb{1}\{x \geq 0\}$ ($\mathbb{1}$ represents the indicator function) to process the output of each of the units. Clearly specify the interconnections and the associated weights. Also argue briefly why your construction is correct.

6. Let X and Y be Boolean valued random variables. Consider the entropy function $H(Y) = \sum_y -P(y)\log(P(y))$ as defined in class. Further, let $H(Y|X) = \sum_x P(x)H(Y|X = x)$ denote the conditional entropy of Y given X, again as defined in class.

   (a) Show that $H(Y) \geq H(Y|X)$.

   (b) Show that $H(Y) = H(Y|X)$ iff Y is independent of X.

   Hint: You can use Jensen's inequality: Let $p(X)$ be a distribution defined over a discrete valued random variable X. Let $f(X)$ be a convex function over X. Then, $E[f(X)] \geq f(E[X])$. Further $E[f(X)] = f(E[X])$ iff $f(X)$ is a constant function.

7. Let $K_1(x, z)$ and $K_2(x, z)$ be two valid kernel functions over $x, z \in \mathcal{R}^n$. Assume that K_1 and K_2 correspond to finite dimensional feature mappings ϕ_1 and ϕ_2, respectively. Show that the function $K(x, z) = K_1(x, z) * K_2(x, z)$ is also a valid kernel.

8. 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 $\sum_{k=1}^{r} n_k = n$. Consider finding the maximum-likelihood estimate of the parameters by maximizing the log-likelihood $LL(\mathbf{p})$ of the data. Note that the data here corresponds to the number of times each face shows up in n trials.

   (a) Set it up as a convex (constrained) optimization problem (think about what the constraints might be on the parameters). Clearly state your objective function, the constraints and the variables you are optimizing over. Show that your formulation is in fact a convex problem. You can use the fact that a linear combination of convex functions is also convex.

   (b) Use the theory of Lagrangians to solve the optimization problem in the part above. Specifically, write the Lagrangian and describe the dual of the problem. Solve the dual problem and convert the solution back to the primal problem.

9. (a) In neural network learning, we initialize the weights close to 0 before starting gradient descent (as opposed to initializing with very high weights). Explain the rationale behind this.

   (b) In decision tree learning, we start with an empty tree and grow it gradually until we hit a desired level of accuracy (as opposed to growing it all the way to the leaf nodes). Therefore, there is an inherent preference over smaller trees. Explain the rationale behind this.

   There is a general principle behind the answer to both of the above questions. Understanding this general idea, complete the following sentence: We start from a _____ hypothesis and gradually _____ it based on the _____. This helps us avoid _____.

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

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \tag{1a}$$

$$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i; \quad \forall i \tag{1b}$$

$$\xi_i \geq 0; \quad \forall i \tag{1c}$$

$$\tag{1d}$$

   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.

   (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.

11. Implement a neural network to represent parity over two Boolean variables, i.e. $y = x_1 \oplus x_2$. Your network should have at most one hidden layer and 3 perceptron units. Use step function as the thresholding units. 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 two real valued inputs. Draw the decision surface enforced by your network in the x_1-x_2 plane.

12. 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.