CIS 520, Machine Learning, Fall 2015: Assignment 1
Due: Friday, September 11th, 11:59pm, PDF to Canvas

Instructions. Please write up your responses to the following problems clearly and concisely. We encourage you to write up your responses using LaTeX; we have provided a LaTeX template, available on Canvas, to make this easier. **Submit your answers in PDF form to Canvas. We will not accept paper copies of the homework.**

Collaboration. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups up to size **two students.** However, *each student must write down the solution independently, and without referring to written notes from the joint session.* **In addition, each student must write on the problem set the names of the people with whom you collaborated.** You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

1 High Dimensional Hi-Jinx [21 points]

Nearest neighbor classifiers can be very flexible and accurate, but what if we have many irrelevant features? Suppose that we have two classes, $y = 1, 2$ and $P(\mathbf{x}|y)$ is Gaussian. Let's consider what happens to distances between points in the same class and in different classes as the dimension of $\mathbf{x}$ grows but only one of the dimensions is informative. We'll assume for simplicity that for all Gaussians below the variance is the same, σ^2. Reminder: If $X \sim N(\mu, \sigma^2)$, $\mathbf{E}[X] = \mu$, $Var[X] = \sigma^2$, and $\mathbf{E}[X^2] = \mu^2 + \sigma^2$. Also, recall that expectation is linear, so it obeys the following three properties:

$$\begin{aligned} \mathbf{E}[X + c] &= E[X] + c \quad \text{for any constant } c \\ \mathbf{E}[X + Y] &= E[X] + E[Y] \\ \mathbf{E}[aX] &= a\mathbf{E}[X] \quad \text{for any constant } a \end{aligned}$$

Lastly, note that if X and X' are independent, then $\mathbf{E}[XX'] = \mathbf{E}[X]\mathbf{E}[X']$.

1. **[4 points]** (Intra-class distance) Consider two points from class 1 in one dimension: $X \sim N(\mu_1, \sigma^2)$ and $X' \sim N(\mu_1, \sigma^2)$. What is the expected squared distance between them? ($\mathbf{E}[(X - X')^2] = \ldots$)

2. **[4 points]** (Inter-class distance) Now consider two points from different classes in one dimension. Now X and X' have different means: $X \sim N(\mu_1, \sigma^2)$ and $X' \sim N(\mu_2, \sigma^2)$. What is the expected squared distance between them? ($\mathbf{E}[(X - X')^2] = \ldots$)

3. **[4 points]** (Intra-class distance, m-dimensions) Again, consider two points from class 1 but in m dimensions. For each dimension j, we have a different mean μ_{1j}: jth dimension of X is $X_j \sim N(\mu_{1j}, \sigma^2)$ and jth dimension of X' is $X'_j \sim N(\mu_{1j}, \sigma^2)$. What is the expected squared distance between them? ($\mathbf{E}[\sum_{j=1}^{m}(X_j - X'_j)^2] = \ldots$)

4. **[4 points]** (Inter-class distance, m-dimensions) Finally, consider two points from different classes but in m dimensions. That is, jth dimension of X is $X_j \sim N(\mu_{1j}, \sigma^2)$ and jth dimension of X' is $X'_j \sim N(\mu_{2j}, \sigma^2)$. What is the expected squared distance between them? ($\mathbf{E}[\sum_{j=1}^{m}(X_j - X'_j)^2] = \ldots$)

5. **[5 points]** Suppose that only one dimension is informative about class values, that is $\mu_{11} \neq \mu_{21}$, but all others have the same mean $\mu_{1j} = \mu_{2j}$ for $j = 2, \ldots, m$. Write down the ratio of expected intra-class distance divided by inter-class distance under this assumption. As m increases towards ∞, what value does this ratio approach?

2 Fitting distributions with KL divergence [28 points]

In this problem, you will use *Kullback-Leibler divergence* (KL-divergence) to measure the difference between two probability distributions. KL-divergence is an important concept in information theory and machine learning. For more on these concepts, refer to Section 1.6 in Bishop. Please note that for most purposes (as is the case for the remainder of the course) the log is base 2.

The KL-divergence from a distribution $p(x)$ to a distribution $q(x)$ can be thought of as a distance measure from P to Q (this is just for intuition, though you should check why this is not a formal distance metric):

$$\begin{aligned} KL(p(x)||q(x)) &= \mathbf{E}_p\left[\log \frac{p(x)}{q(x)}\right] \\ &= \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x && \text{For continuous } p \text{ and } q \\ &= \sum p(x) \log \frac{p(x)}{q(x)} && \text{For discrete } p \text{ and } q \end{aligned}$$

From an information theory perspective, the KL-divergence specifies the number of additional bits required on average to transmit values of x if the values are distributed with respect to $p(x)$ but we encode them assuming the distribution $q(x)$. If $p(x) = q(x)$, then $KL(p||q) = 0$. Otherwise, $KL(p||q) > 0$. The smaller the KL-divergence, the more similar the two distributions.

1. (a) **[6 points]** Provide the formula the Kullback-Leibler divergence $KL(p(x)||q(x))$ between two univariate Gaussians distributions:

$$p(x) = \mathcal{N}(\mu_1, \sigma^2), \qquad q(x) = \mathcal{N}(\mu_2, 1).$$

   Write your answer in terms of expectation over p; i.e., fill in f and g such that $KL(p(x)||q(x)) = \mathbf{E}_p[f(x, \mu_1, \mu_2, \sigma)] + g(\sigma)$.

   (b) **[7 points]** For fixed μ_2 and σ, what value of μ_1 minimizes $KL(p(x)||q(x))$? At the minimum, what is the value of $KL(p(x)||q(x))$? Your answers should depend only on μ_2 and/or σ.

2. (a) **[3 points]** Let $q(x)$ be a Multinomial distribution over $\{1, \ldots, k\}$, with parameters $\{\theta_1, \ldots, \theta_k\}$ (remember that $\sum_j \theta_j = 1$). Let $p(x)$ over $\{1, \ldots, k\}$ be the following distribution:

$$p(x) = \begin{cases} \alpha & \text{if } x \text{ is even} \\ \beta & \text{if } x \text{ is odd.} \end{cases} \tag{1}$$

   Note that $\sum_{x=1}^{k} p(x) = 1$ for $p(x)$ to be a probability distribution. For simplicity, assume that $k = 2n$, in other words, that k is even. Then we have the constraint that $n(\alpha + \beta) = 1$.

   Provide the formula for KL divergence $KL(p(x)||q(x))$ in terms of α, β, and $\theta_1, \ldots, \theta_k$.

   Hint: Please use the shorthand $\sum_{i \text{ even}}$ to indicate summation over even elements in $\{1, \ldots, k\}$, and similarly for odd elements. Also for statisticians, the multinomial distribution here is a categorical distribution, where k possible outcomes each have associated probabilities. So one needs to specify $k - 1$ probabilities that sum to 1.

   (b) **[12 points]** For fixed $\theta_1, \ldots, \theta_k$, what values of α and β achieve lowest $KL(p(x)||q(x))$? For this problem, You do not need to give the value at the minimum, simply find the best α and β. Your answers should depend only on $\theta_1, \ldots, \theta_k$.

Hint: To solve this problem, you will need to use the method of Lagrange multipliers to incorporate the constraint. I.e., optimize (take partial derivatives and set to zero)

$$KL(p(x)||q(x)) + \lambda(n(\alpha + \beta) - 1),$$

with respect to α, β, and λ together. Do not worry about the implicit constraints $\alpha, \beta \geq 0$; KL will handle these for you.

3 Conditional Independence in Probability Models [16 points]

Consider the following probability model. For a set of points $x_1, \ldots, x_n$, we have k possible generating distributions, $f_1, \ldots, f_k$. (That is, we know each point x_i was generated from one of the f_j; we just don't know which j.) Let $z_i = \{1, \ldots, k\}$ be an indicator variable which indicates that the i'th data point x_i was generated from f_j if $z_i = j$. Furthermore, we specify that $p(z_i = j) = \pi_j$. Thus, our model specifies the following:

$$\text{(i) } p(x_i \mid z_i = j) = f_j(x_i), \qquad \text{(ii) } p(z_i = j) = \pi_j.$$

Hint: The key to this problem is applying the basic rules of probability–marginalization, Bayes Rule, the chain rule, and/or conditional independence (not necessarily in that order). Note that some of the answers may follow directly from the definitions given above.

1. **[4 points]** Derive the formula for $p(x_i)$ in terms of (i) and (ii) above.
2. **[6 points]** Derive the formula for $p(x_1, \ldots, x_n)$ in terms of (i) and (ii) above.
3. **[6 points]** Derive the formula for $p(z_u = v \mid x_1, \ldots, x_n)$ in terms of (i) and (ii) above.

4 Decision Trees [22 points]

1. Consider the following set of training examples for the unknown target function $< X_1, X_2 > \rightarrow Y$. Each row indicates the values observed, and how many times that set of values was observed. For example, $(+, T, T)$ was observed 6 times, while $(-, T, T)$ was observed once.

Y	X_1	X_2	Count
+	T	T	6
+	T	F	0
+	F	T	2
+	F	F	4
-	T	T	1
-	T	F	2
-	F	T	1
-	F	F	4

Table 1:

   (a) **[3 points]** What is the sample entropy $H(Y)$ for this training data (with logarithms base 2)?

   (b) **[6 points]** What are the information gains $IG(X_1) \equiv H(Y) - H(Y|X_1)$ and $IG(X_2) \equiv H(Y) - H(Y|X_2)$ for this sample of training data?

   (c) **[8 points]** Draw the decision tree that would be learned by ID3 (without postpruning) from this sample of training data. (Hint: Feel free to include a photo instead of using latex. Also break ties by voting +)

2. **[5 points]** When we discussed learning decision trees in class, we chose the next attribute to split on by choosing the one with maximum information gain, which was defined in terms of entropy. To further our understanding of information gain, we will explore its connection to KL-divergence (introduced above).

   We can define information gain as the KL-divergence from the observed joint distribution of X and Y to the product of their observed marginals.[1]

$$IG(x,y) \equiv KL\left(p(x,y)||p(x)p(y)\right) = -\sum_x \sum_y p(x,y)\ \log\left(\frac{p(x)p(y)}{p(x,y)}\right)$$

   When the information gain is high, it indicates that adding a split to the decision tree will give a more accurate model.

   Show that this definition of information gain is equivalent to the one given in class. That is, show that $IG(x,y) = H[x] - H[x|y] = H[y] - H[y|x]$, starting from the definition in terms of KL-divergence.

5 Non-Normal Norms [13 points]

1. **[4 points]**

   Given the following four vectors,
   $x_1 = [1.0, 0.5, 0.7, 3.0]$
   $x_2 = [0.6, 2.0, 2.3, 2.1]$
   $x_3 = [1.0, 0.5, 1.4, 10.1]$
   $x_4 = [0.1, 0, 2.7, 3.5]$
   Which point is closest to x_1 under each of the following norms
   a) L_0
   b) L_1
   c) L_2
   d) L_{inf}

2. **[9 points]** You want to build a predictive model using machine learning to estimate the cost of a software project that you are bidding to do. If you underestimate the cost, you will have to cover the extra cost. If you overestimate the cost, you will not get the contract, and make no money. (Sorry, as phrased here, this is a no-win situation, since I am assuming that if you price it perfectly, you will exactly get paid for the value of your labor.)

   (a) **[3 points]** Write a formula for your "loss function" as a function of $y - \hat{y}$

   (b) **[3 points]** Is your loss function a norm?
   (see the formal definition at http://en.wikipedia.org/wiki/Norm_(mathematics)#Definition)

   (c) **[3 points]** Which norm of the form $||y - \hat{y}||_p$ do you think would be best to use to measure error in this case and (in one sentence) why?

[1] The negative sign is introduced in this definition because $\log(p/q) = -\log(q/p)$; flipping the fraction will give us KL as defined previously.