CS 6140: Machine Learning **Samuel Steiner steiner.s@husky.neu.edu**

Homework Assignment #1

Assigned : *01/29/2021* *Due: 02/12/2021, 11:59pm, through Canvas*

Three problems, 130 points in total. Good luck!
Prof. Predrag Radivojac, Northeastern University

Problem 1. (10 points) Let X, Y and Z be discrete random variables defined as functions on the same probability space $(\Omega, \mathcal{A}, P)$. Prove or disprove the following expression

$$P(X = x \mid Y = y) = \sum_{(z \in \mathcal{Z})} P(X = x \mid Y = y, Z = z) \cdot P(Z = z \mid Y = y)$$

where $\mathcal{Z}$ is the sample space defined by the random variable Z

Proof. Lets say P_y represents the conditional probability $P(\cdot \mid Y = y)$.

Using this we can rewrite the statement as follows

$$P_y(X = x) = \sum_{(z \in \mathcal{Z})} P_y(X = x \mid Z = z) \cdot P_y(Z = z)$$

Since $z \in \mathcal{Z}$ This is by definition of the law of total probability.

Which can be proven as follows:

$$\begin{aligned} P(X = x) &= P(X = x \cap \bigcup_{z \in \mathcal{Z}} Z = z) \text{ Where } \mathcal{Z} \text{ is the sample space of Z} \\ &= P(\bigcup_{z \in \mathcal{Z}} X = x \cap Z = z) \\ &= \sum_{z \in \mathcal{Z}} P(X = x, Z = z) \\ &= \sum_{z \in \mathcal{Z}} P(X = x \mid Z = z) P(Z = z) \end{aligned}$$

□

Problem 2. (15 points) Let X be a random variable on $\mathcal{X} = \{a, b, c\}$ with the probability mass function $p(x)$. Let $p(a) = 0.1, p(b) = 0.2$, and $p(c) = 0.7$ and some function $f(x)$ be

$$f(x) = \begin{cases} 10 & x = a \\ 5 & x = b \\ \frac{10}{7} & x = c \end{cases}$$

a) (5 points) What is $\mathbb{E}[f(X)]$?

$$\mathbb{E}[f(X)] = f(a) * 0.1 + f(b) * 0.2 + f(c) * 0.7 = 1 + 1 + 1 = 3$$

b) (5 points) What is $\mathbb{E}[1/p(X)]$?

$$\mathbb{E}[1/p(x)] = (1/p(a)) * p(a) + (1/p(b)) * p(b) + (1/p(c)) * p(c) = 1 + 1 + 1 = 3$$

c) (5 points) For an arbitrary finite set $\mathcal{X}$ with n elements and arbitrary $p(x)$ on $\mathcal{X}$ what is $\mathbb{E}[1/p(X)]$?

The value of this should always be n since by definition $\mathbb{E}[K] = \sum_{k \in K} (kp(k))$ in this case K is replaced with $1/p(X)$ which gets us $\mathbb{E}[1/p(X)] = \sum_{x \in X} 1/p(x) * p(x) = \sum_{x \in X} 1$ which is equal to the size of X or n.

Problem 3. (15 points) A biased four sided die is rolled down and the down face is a random variable X described by the following pmf.

$$p(x) = \begin{cases} x/10 & x = 1, 2, 3, 4 \\ 0 & \text{otherwise} \end{cases}$$

Given the random variable X a biased coin is flipped and the random variable Y is 1 or zero according to whether the coin shows heads or tails. The conditional pmf is

$$p(y \mid x) = \left(\frac{x+1}{2x}\right)^y \left(1 - \frac{x+1}{2x}\right)^{1-y}$$

Where $y \in \{0,1\}$.

a) (5 points) Find the expectation $\mathbb{E}[X]$ and the variance $V[X]$

$$\begin{aligned}\mathbb{E}[X] &= 1(\frac{1}{10}) + 2(\frac{2}{10}) + 3(\frac{3}{10}) + 4(\frac{4}{10}) \\ &= 3\end{aligned}$$

$$\begin{aligned}V(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= (1(\frac{1}{10}) + 4(\frac{2}{10}) + 9(\frac{3}{10}) + 16(\frac{4}{10})) - 3^2 \\ &= \frac{1}{10} + \frac{8}{10} + \frac{27}{10} + \frac{64}{10} - 9 \\ &= 1\end{aligned}$$

b) (5 points) Find the conditional pmf $p(x \mid y)$

$$\begin{aligned}p(y) &= \sum_{x \in X} p(y|x)p(x) \\ &= \sum_{x=1}^{4} \left(\frac{x+1}{2x}\right)^y \left(1 - \frac{x+1}{2x}\right)^{1-y} \left(\frac{x}{10}\right) \\ &= \frac{(2 \cdot 0^{(1-y)} + 2^{(1+y)} + 3^y + 3^{(1-y)} \cdot 5^y)}{20} \\ &\text{since } y \in \{0,1\} \text{ we can write this simply as} \\ &= \left(\frac{7}{10}\right)^y \left(\frac{3}{10}\right)^{1-y} \\ p(x \mid y) &= \frac{p(x)p(y \mid x)}{p(y)} \\ &= \frac{\frac{x}{10}\left(\frac{x+1}{2x}\right)^y \left(1 - \frac{x+1}{2x}\right)^{1-y}}{\left(\frac{7}{10}\right)^y \left(\frac{3}{10}\right)^{1-y}} \\ &\text{Which can be written as} \\ &= \begin{cases} \frac{\frac{x}{10}\left(\frac{x+1}{2x}\right)}{\frac{7}{10}} & y = 1 \\ \frac{\frac{x}{10}\left(1 - \frac{x+1}{2x}\right)}{\frac{3}{10}} & y = 0 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

c) (5 points) Find the conditional expectation $\mathbb{E}[X \mid Y = 1]$; i.e., the expectation with respect to the conditional pmf $p_{X|Y}(x \mid 1)$.

$$\begin{aligned}\mathbb{E}[X \mid Y = 1] &= \sum_{x=1}^{4} x \left(\frac{\frac{x}{10}\left(\frac{x+1}{2x}\right)}{\frac{7}{10}}\right) \\ &= \frac{20}{7}\end{aligned}$$

Problem 4. (25 points) suppose that data set $\mathcal{D} = \{1,0,1,1,1,0,1,1,1,0\}$ is an i.i.d. sample from a Bernoulli distribution

$$p(x \mid \alpha) = \alpha^x (1-\alpha)^{1-x} \qquad 0 < \alpha < 1$$

with unknown parameter α

a) (5 points) Calculate the log-likelihood of the data $\mathcal{D}$ when $\alpha = \frac{1}{e}$; i.e., find $\log p(\mathcal{D} \mid \alpha = 1/e)$. The parameter e is the Euler number. Write the final expression as compactly as you can.

$$\sum_{d \in \mathcal{D}} \log p(d \mid \alpha) = \sum_{d \in \mathcal{D}} \log\left((1/e)^d (1-(1/e))^{1-d}\right)$$

Which maps

$$= \begin{cases} \log(1/e) = -1 & 1 \\ \log(1-(1/e)) \approx -0.46 & 0 \end{cases}$$

for the data set $\mathcal{D}$

$$= -7 + 3(\log(1-(1/e)))$$

$$= -10 + 3(\log(e-1))$$

b (10 points) Compute the maximum likelihood estimate of α. Show all your work.

$$\hat{\alpha} = \arg\max_{\alpha} LL(\alpha)$$

$$LL(\alpha) = \sum_{d \in \mathcal{D}} \log(((\alpha)^d (1-(\alpha))^{1-d})$$

with our dataset this becomes

$$= 7(\log \alpha) + 3(\log(1-\alpha))$$

now we do the first derivative and set it to 0

$$\frac{\delta LL(\alpha)}{\delta \alpha} = \frac{7}{\alpha} + \frac{3}{\alpha - 1} = 0$$

$$0 = \frac{10\alpha - 7}{(\alpha-1)\alpha}$$

$$0 = 10\alpha - 7$$

$$7 = 10\alpha$$

$$\frac{7}{10} = \alpha$$

c (10 points) Suppose the prior distribution for α is the uniform distribution on $(0,1)$ compute the Bayes estimator for α. Note that $\int_0^1 v^m (1-v)^r dv = \frac{m!r!}{(m+r+1)!}$.

$$\arg\max_{\alpha} p(\alpha \mid \mathcal{D}) = p(\mathcal{D} \mid \alpha) p(\alpha)$$

$$= LL(p(\mathcal{D} \mid \alpha) p(\alpha))$$

$$= \log(p(\mathcal{D} \mid \alpha)) + \log p(\alpha)$$

$$= \log(p(\mathcal{D} \mid \alpha)) + \log p(1)$$

We already solved maximization of$\log p(\mathcal{D} \mid \alpha)$

$$0 = 10\alpha - 7$$

$$7 = 10\alpha$$

$$\frac{7}{10} = \alpha$$

Problem 5. (10 points) Let $\mathcal{D} = \{X_i\}_{i=1}^n$ be an i.i.d. sample from

$$p(x) = \begin{cases} e^{-(x-\theta_0)} & x \geq \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

Determine θ_{ML} - the maximum likelihood estimate of θ_0.

$$\begin{aligned}\log p(\mathcal{D} \mid \theta_0) &= \log \prod_{i=1}^{n} p(x_i \mid \theta_0) \\ &= \sum_{i=1}^{n} \log e^{-(x-\theta_0)} \\ &\text{Now take the derivative}\end{aligned}$$

$$\frac{\delta}{\delta\theta_0} \sum_{i=1}^{n} \log e^{-(x-\theta_0)} = n$$

Problem 6. (25 Points) Understanding the curse of dimensionality. Consider the following experiment: generate n data points with dimensionality k. Let each data point be generated using a uniform random number generator with values between 0 and 1. Now, for a given k, calculate

$$r(k) = \log_{10} \frac{d_{\max}(k) - d_{\min}(k)}{d_{\text{ave}}(k)}$$

where $d_{\max}(k)$ is the maximum distance between any par of points, $d_{\min}(k)$ is the minimum distance between any pair of points (you cannot use identical points to obtain the minimum distance of 0), and d_{ave} is the average distance between pairs of distinct points in the data set. Let k take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each k.

a) (10 points) Using Euclidean distance to compute $d_{\max}$ and $d_{\min}$, plot $r(k)$ as a function of k for two different values of $n; n \in \{100, 1000\}$. Label and scale each axis properly to be able to make comparisons over different n's. Embed your final picture(s) in the file you are submitting for this assignment.

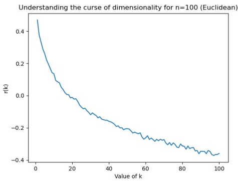

Figure 1: Curse of dimensionality $n = 100$
Euclidean Distance

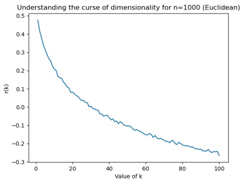

Figure 2: Curse of dimensionality $n = 1000$
Euclidean Distance

b) (10 points) Replace Euclidean distance by the cosine distance, defined as $d_{\cos}(x, y) = 1 - \cos(x, y)$, where x and y are k-dimensional data points and $\cos(x, y)$ is the cosine similarity. Then repeat the experiment from part a.

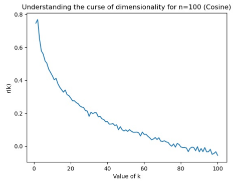

Figure 3: Curse of dimensionality $n = 100$
Cosine Distance

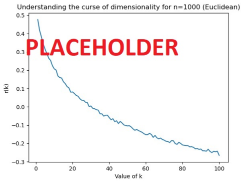

Figure 4: Curse of dimensionality $n = 1000$
Cosine Distance

c) (5 points) Discuss your observations and also compare the results to your expectations before you carried out the experiments in parts a and b.

My initial thoughts were that the graphs would like a normal log graph where the denominator is growing larger than the numerator that is to say the graph trends in the negative direction and 'flattens' out. This is all I really assumed when I first saw the experiment. As I predicted this is what the result was, this shows that with higher dimensionality it may be harder to find distinct differences in the data. The 1000 cosine, graphic was corrupted and I was unable to reproduce it by the time submission was necessary.

Problem 7. (30 points) Expectation-Maximization (EM) algorithm. Let X be a random variable distributed according to

$$p(x) = \alpha q(x \mid \lambda_1) + (1-\alpha) q(x \mid \lambda_0)$$

where $\alpha \in (0,1)$ Let $q(x \mid \lambda) = \frac{\lambda}{x^{\lambda+1}}$ on the input space $[1, \inf)$ be a Pareto distribution with $\lambda > 0$. Let now $\mathcal{D} = \{x_i\}_{i=1}^n$ be a set of observations.

a) (10 points) Derive update rules of the EM algorithm to estimate, α, λ_0, and λ_1.

$$\begin{aligned} \theta^t &= (\alpha^t, \lambda_0^t, \lambda_1^t) \\ \alpha_0 &= \alpha \\ k &= \{0,1\} \\ \text{1.intialize } \alpha, \lambda_0, &\text{and } \lambda_1 \text{at } t = 0 \\ \text{2.Repeat until convergence} & \\ \text{(a) } q_{Y_i}(k \mid x_i, \theta^t) &= \frac{\alpha_k^t q(x_i | \lambda_k^t)}{\sum_{j=0}^1 \alpha_j^t q(x_i \mid \lambda_j^t)} \text{for } \forall (i,k) \\ \text{(b) } \alpha_k^{t+1} &= \frac{1}{n} q_{Y_i}(k \mid x_i, \theta^t) \\ \text{(c) } \lambda_k^{t+1} &= \frac{\sum_{i=1}^n q_{Y_i}(k \mid x_i, \lambda^t)}{\sum_{i=1}^n x_i q_{Y_i}(k \mid x_i, \theta^t)} \\ \text{(d) } t &= t+1 \end{aligned}$$

b) (20 points) Implement the learning algorithm from part a and evaluate it on 100 simulated datasets with n no less than 100. Each dataset should be generated according to a distribution with fixed parameters. To assess the quality of your estimates, visualize the distribution of absolute differences between estimated and true parameters using box plots and compute the mean absolute difference. Discuss your experiments, discuss steps and calls you needed to make, and report on the quality of your algorithm.