CS6140: Machine Learning

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

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

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

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 $\mathbb{E}[f(X)]$?

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

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

Problem 3. (15 points) A biased four-sided die is rolled 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

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

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

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

b) (5 points) Find the conditional pmf $p(x|y)$.

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

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

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

with an unknown parameter α.

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

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

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

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.

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 pair of points, $d_{\min}(k)$ is minimum distance between any pair of points (you cannot use identical points to obtain the minimum distance of 0), and $d_{\text{ave}}(k)$ 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.

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 cosine similarity. Then repeat the experiment from part a.

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.

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

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

where $\alpha \in (0, 1)$, Let $q(x|\lambda) = \frac{\lambda}{x^{\lambda+1}}$ on the input space $[1, \infty)$ 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.

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.

Directions and Policies

Submit a single package containing all answers, results and code. Your submission package should be compressed and named firstnamelastname.zip (e.g., predragradivojac.zip). In your package there should be a single pdf file named main.pdf that will contain answers to all questions, all figures, and all relevant results. Your solutions and answers must be typed[1] and make sure that you type your name and Northeastern username (email) on top of the first page of the main.pdf file. The rest of the package should contain all code that you used. The code should be properly organized in folders and subfolders, one for each question or problem. All code, if applicable, should be turned in when you submit your assignment as it may be necessary to demo your programs to the teaching assistants. Use Matlab, Python, R, Java, or C/C++. However, you are encouraged to use languages with good machine learning libraries (e.g., Matlab, Python, R), which may be handy in future assignments.

Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rules:

on time: your score $\times$ 1

1 day late: your score $\times$ 0.9

2 days late: your score $\times$ 0.7

3 days late: your score $\times$ 0.5

4 days late: your score $\times$ 0.3

5 days late: your score $\times$ 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. **All the sources used for problem solution must be acknowledged**; e.g., web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously! For detailed information see Office of Student Conduct and Conflict Resolution.

[1] We recommend Latex; in particular, TexShop-MacTeX combination for a Mac and TeXnicCenter-MiKTex combination on Windows. An easy way to start with Latex is to use the freely available Lyx. You can also use Microsoft Word or other programs that can display formulas professionally.