CS6140: Machine Learning

Homework Assignment # 1

Assigned: 09/14/2024 *Due: 09/30/2024, 11:59pm, through Canvas*

Eight problems, 170 points in total. Good luck!
Prof. Predrag Radivojac, Northeastern University

Problem 1. (20 points) Two players perform a series of coin tosses. Player one wins a toss if the coin turns heads and player two wins a toss if it turns tails. The game is played until one player wins n times. However, the game is interrupted when player one had l wins and player two had m wins, where $0 \leq l, m < n$.

a) (5 points) Assuming $n = 8$, $l = 4$, and $m = 6$, what is the probability that player one would win the game if the game was to be continued later.

b) (10 points) Derive the general expression or write an algorithm that computes the probability that player one will win the game if the game is to be continued later. Your expression should be a function of n, m, and l. If you are providing an algorithm, implement it and submit your code along with your pseudo-code that should be in your report. You may *not* simulate the game as your solution, though you can use simulation to verify your solution. Hint: negative binomial distribution may be useful.

c) (5 points) When l and m are kept constant, describe the influence of n to the final probability. How does that compare to what you thought about the problem before you solved it? Was your intuition right?

Problem 2. (5 points) Let $(\Omega, \mathcal{A}, P)$ be a discrete probability space, where $\mathcal{A} = \mathcal{P}(\Omega)$, and let $A \subseteq \Omega$ and $B \subseteq \Omega$ be any two subsets of Ω. Prove the following expression or provide a counterexample if it does not hold

$$P(A) = P(A|B) + P(A|B^c),$$

where A^c is the complement of A.

Problem 3. (25 points) Mary is going shopping for books to take on a trip. She will spend X hours in the bookstore, where X is a discrete random variable equally likely to take on values of 1, 2, 3, or 4. She will buy Y books, where Y is a discrete random variable that depends on the amount of time she shops as described by a conditional probability mass function

$$p(y|x) = \frac{1}{x}; \quad y = 1, 2, \ldots, x$$

a) (5 points) Find the joint probability mass function of X and Y.

b) (5 points) Find the marginal probability mass function for Y.

c) (5 points) Find the conditional probability mass function of X given that $Y = 2$

d) (5 points) Suppose you know that Mary bought at least two but not more than three books. Find the conditional mean and variance of X given this information.

e) (5 points) The cost of each book is a random variable (independent of all the other random variables mentioned) with mean 3. What is the total expected expenditure of Mary's visit to the bookstore?

Problem 4. (15 points) Let Y_0 and Y_1 be two continuous random variables and $Z \sim \text{Bernoulli}(\alpha)$. Let X be a random variable defined as $X = ZY_1 + (1 - Z)Y_0$. Assuming that the probability density functions of Y_0 and Y_1 exist, show that the density of X is a mixture of the densities of Y_1 and Y_0 with α and $1 - \alpha$ as the mixing proportions, respectively.

Problem 5. (20 points) Suppose that the number of accidents occurring daily in a certain plant has a Poisson distribution with an unknown mean λ. Based on previous experience in similar industrial plants, suppose that our initial feelings about the possible value of λ can be expressed by an exponential distribution with parameter $\theta = \frac{1}{2}$ is, the prior density is

$$p(\lambda) = \theta e^{-\theta\lambda}$$

where $\lambda \in (0, \infty)$. If there are 72 accidents over the next 8 days, determine

a) (5 points) the maximum likelihood estimate of λ

b) (5 points) the maximum a posteriori estimate of λ

c) (10 points) the Bayes estimate of λ.

Problem 6. (15 points) Let X_1, X_2, ..., X_n be i.i.d. Gaussian random variables, each having an unknown mean θ and known variance σ_0^2. If θ is itself selected from a normal population having a known mean μ and a known variance σ^2, determine

a) (5 points) the maximum a posteriori estimate of θ

b) (10 points) the Bayes estimate of θ.

Hint: look into conjugate priors for the Gaussian distribution. Chapter 2 of Bishop's textbook will be useful as well as resources on the internet.

Problem 7. (40 points) Expectation-maximization (EM) algorithm. Let X be a random variable distributed according to $p_X(x)$ and Y be a random variable distributed according to $p_Y(y)$. Let $\mathcal{D}_X = \{x_i\}_{i=1}^m$ be an i.i.d sample from $p_X(x)$ and $\mathcal{D}_Y = \{y_i\}_{i=1}^n$ be an i.i.d. sample from $p_Y(y)$. Finally, let $p_X(x)$ and $p_Y(y)$ be defined as follows

$$p_X(x) = \alpha\mathcal{N}(\mu_1, \sigma_1^2) + (1 - \alpha)\mathcal{N}(\mu_2, \sigma_2^2)$$

and

$$p_Y(y) = \beta\mathcal{N}(\mu_1, \sigma_1^2) + (1 - \beta)\mathcal{N}(\mu_2, \sigma_2^2),$$

where $\mathcal{N}(\mu, \sigma^2)$ is a univariate Gaussian distribution with mean μ and variance σ^2, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$, $\sigma_1 \in \mathbb{R}^+$ and $\sigma_2 \in \mathbb{R}^+$ are unknown parameters.

a) (5 points) Derive update rules of the EM algorithm for estimating α, μ_1, μ_2, σ_1, and σ_2 based only on data set $\mathcal{D}_X$.

b) (15 points) Derive update rules of an EM algorithm for estimating α, β, μ_1, μ_2, σ_1, and σ_2 from data sets $\mathcal{D}_X$ and $\mathcal{D}_Y$.

c) (20 points) Implement both learning algorithms from above and evaluate them on simulated data when $m, n = 100$ and $m, n = 1000$. However, in each case repeat the experiment $B = 1000$ times to estimate the expectation and variance of all parameters. To do so, repeatedly draw samples $\mathcal{D}_X$ and $\mathcal{D}_Y$ and then estimate the parameters based on these samples. Finally, average those B estimates and calculate their mean and variance. Document all experiments and discuss your findings.

To implement the EM algorithm you must make some practical decisions on how to stop the estimation process. You may decide to impose the maximum number of steps, stop the algorithm when the updated parameters stabilize, stop the algorithm when the log-likelihood stabilizes, or some combination thereof. In either case, experiment first with the stoppage criterion and once it is fixed carry out the experiment.

Problem 8. (30 points) Properties of high-dimensional spaces.

a) (10 points) Show that in a high-dimensional space, most of the volume of a hypercube is concentrated in corners, which themselves become very long "spikes." Hint: compute the ratio of the volume of a hypersphere with radius r to the volume of a hypercube with side $2r$ around it and also the ratio of the distance from the center of the hypercube to one of the corners divided by the distance to one of the sides.

b) (10 points) Show that for points uniformly distributed inside a sphere in d dimensions, where d is large, almost all of the points are concentrated in a thin shell close to the surface. Hint: compute the fraction of the volume of the sphere which lies at the distance between $r - \epsilon$ and r from the center, where $0 < \epsilon < r$. Evaluate this fraction for $\epsilon = 0.01r$ and also for $\epsilon = 0.5r$ for $d \in \{1, 2, 3, 10, 100\}$.

c) (10 points) Evaluate computationally what you derived. First, generate n d-dimensional data points uniformly at random within a hypercube with side $2r$. Then, compute the fraction of points f that are within the hypersphere of radius r inscribed in the hypercube. Do this for d ranging from 1 to 100. Generate the plot of f as a function of d (make sure axes are labeled). Pick n to be at least 100, but a larger number is desirable, depending on your computational resources and also mark the values computed from the formula derived in part (a). If you pick a relatively small n, and have computational resources, visualize the uncertainty of the estimated fraction over multiple trials for the same d. Evaluate also what happens if you vary r for a fixed d? Describe what you see in all plots and give your reasoning as to why it happened.

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 or R.

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 text must be be your own or, for group assignments (when explicitly permitted), by the members of the group. All 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.