Homework 1 - Probability and Priors

DS6040 Bayesian Machine Learning - Fall 2024
Instructor: Teague R. Henry (ycp6wm@virginia.edu)

This assignment is 100 points in total with 25 additional points available for extra credit.

Problem 1: Basic Probability
Alice has a bag with 3 red balls, 2 green balls, and 5 blue balls.

(a) What is the probability of drawing a red ball from the bag?

(b) If Alice draws one ball and it's blue, what's the probability that the next ball she draws is also blue?

Problem 2: Independent Events
The probability of a server being down in a data center is 0.05. The data center is designed such that server failures are independent events.

(a) What is the probability that 2 servers will be down at the same time?

(b) What is the probability that at least one of two servers will be down?

Problem 3: Conditional Probability
In a Machine Learning company, 30% of the employees are Data Scientists, 40% of the Data Scientists have PhDs, while only 10% of non-Data Scientists have PhDs.

(a) If an employee is chosen randomly, what is the probability that the employee is a Data Scientist with a PhD?

(b) Given that an employee has a PhD, what is the probability that the employee is a Data Scientist?

Problem 4: Law of Total Probability
A diagnostics test has a probability of 0.95 of giving a positive result when applied to a person suffering from a certain disease. It has a probability of 0.10 of giving a (false) positive result when applied to a non-sufferer. It is estimated that 0.5% of the population has this disease.

(a) If a person tested positive in the test, what is the probability that the person actually has the disease?

(b) What is the total probability of a person testing positive?

Problem 5: Bayes' Theorem
An email filter is set up to classify emails into "spam" and "not spam". It is known that 90% of all emails received are spam. The filter correctly identifies spam 95% of the time and correctly identifies "not spam" 85% of the time.

(a) If an email is picked at random, and the filter classifies it as spam, what is the probability that it is actually spam?

(b) If an email is classified as "not spam", what is the probability that it is actually spam?

Problem 6: Expectation of a Discrete Random Variable
Consider a dice game where you roll a fair six-sided die. If a 6 appears, you win $10. If any other number appears, you lose $2.

(a) Define the random variable X that models this game.

(b) Compute the expected value of X.

Problem 7: Expectation of a Continuous Random Variable
Let X be a continuous random variable representing the time (in hours) it takes for a server to process a certain type of query. Suppose the density function of X is given by $f(x) = 2e^{-2x}$ for $x \geq 0$.

(a) Compute the expected value $E[X]$ of X.

(b) Compute the variance $Var[X]$ of X.

(c) Interpret your findings from parts (a) and (b) in the context of the server's processing time.

Problem 8: Markov Chain
Consider a simple weather model defined by a Markov chain. The weather on any given day can be either "sunny", "cloudy", or "rainy". The transition probabilities are as follows:

- If it is sunny today, the probabilities for tomorrow are: 0.7 for sunny, 0.2 for cloudy, and 0.1 for rainy.
- If it is cloudy today, the probabilities for tomorrow are: 0.3 for sunny, 0.4 for cloudy, and 0.3 for rainy.
- If it is rainy today, the probabilities for tomorrow are: 0.2 for sunny, 0.3 for cloudy, and 0.5 for rainy.

A *transition matrix* is a square matrix describing the transitions of a Markov chain. Each row of the matrix corresponds to a current state, and each column corresponds to a future state. Each entry in the matrix is a probability.

(a) Construct the transition matrix for this Markov chain.

After many iterations or steps, the probabilities of being in each state may stabilize to a constant value. These constant values form the *stationary distribution* of the Markov chain. To compute the stationary distribution, find the probability vector that remains unchanged after multiplication with the transition matrix.

(b) If today is sunny, what is the probability that it will be rainy two days from now?

(c) Find the stationary distribution of this Markov chain.

(d) Interpret the stationary distribution in the context of this weather model.

Problem 9: Conjugate Priors and Posterior Distribution

In Bayesian inference, the Beta distribution serves as a conjugate prior distribution for the Bernoulli, binomial, negative binomial, and geometric distributions. For a single observed data point, the Bernoulli distribution can be written as:

$P(x|\theta) = \theta^x \cdot (1-\theta)^{1-x}$

where $x \in \{0, 1\}$ and $0 \leq \theta \leq 1$.

The beta distribution is a suitable conjugate prior for θ. It's given by:

$P(\theta|\alpha, \beta) = \frac{\theta^{(\alpha-1)} \cdot (1-\theta)^{(\beta-1)}}{B(\alpha,\beta)}$

where $B(\alpha, \beta)$ is the beta function, and α and β are the parameters of the beta distribution.

Now consider an experiment where a new drug is tested on 100 patients. Out of these, 30 patients recover.

(a) Suppose the prior distribution for θ (the recovery rate) is Beta(2, 2). Calculate the posterior distribution after observing the results of the experiment.

(b) Based on the posterior distribution, provide an estimate for θ.

(c) Explain the role of the conjugate prior in simplifying the calculation of the posterior distribution.

Extra Credit: Non-Informative Priors (25 points)

In Bayesian inference, when little is known about the prior distribution, non-informative priors are often used. Two common types of non-informative priors are conjugate and Jeffreys priors.

Consider a model where we have a normal likelihood with known standard deviation ($\sigma = 5$), and an unknown mean (μ), which we're trying to infer from data. We have a dataset X = $\{x_1$, x_2, ..., $x_n\}$, drawn independently from this normal distribution.

(a) Suppose we use a vague (non-informative) conjugate prior for μ, i.e., a normal distribution $N(0, 100^2)$. Compute the posterior distribution for μ after observing the data.

(b) Now consider using a Jeffreys prior. For a normal distribution with known σ and unknown μ, the Jeffreys prior is a improper flat (uniform) prior, which implies that every possible value of μ is equally likely a priori. This can be represented as:

$P(\mu) = 1$, for $-\infty < \mu < \infty$

Compute the posterior distribution for μ in this case.

(c) Compare the posterior distributions from the vague conjugate prior and the Jeffreys prior. How do they differ and what might cause this difference?

(d) Discuss the pros and cons of using non-informative priors in general, and how choosing different types of non-informative priors might affect your inferences.