Matt Scheffel – mcs9ff
DS 6040
6/12/23

Homework 1 – Probability and Priors

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?
   a. The probability of drawing a red ball from the bag is:
      i. P(R) = 3/10 = 30%

(b) If Alice draws one ball and it's blue, what's the probability that the next ball she draws is also blue?
   a. The probability that the next ball she draws is blue is:
      i. P(B) = 4/9 = 44.4%

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?
   a. P(2 Down) = P(1 Down) x P(1 Down) = 0.05 x 0.05 = 0.0025 = 0.25%
   b. The probability that 2 servers are down at the same time is 0.25%.

(b) What is the probability that at least one of two servers will be down?
   a. P(1 Up) = 1 - P(1 Down) = 1 - 0.05 = 0.95
   b. P(Both Up) = P(1 Up) x P(1 Up) = 0.95 x 0.95 = 0.9025
   c. P(At Least 1 Down) = 1 – P(Both Up) = 1 – 0.9025 = .0975 = 9.75%
   d. The probability that at least 1 server is down is 9.75%.

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?
   a. P(Data Scientist w/ PhD) = P(Data Scientist) × P(PhD | Data Scientist)
   b. P(Data Scientist w/ PhD) = 0.30 x 0.40 = 0.12 = 12%
   c. If an employee is chosen randomly, the probability that the employee is a Data Scientist with a PhD is 12%.

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

a. Use Bayes' Theorem: P(A|B) = (P(B|A) x P(A)) / P(B)
    i. With A = Data Scientist and B = PhD
        1. We know P(B|A) = 0.40
        2. We know P(A) = 0.30
    ii. P(B) = P(PhD) = P(B|A) x P(A) + P(B|A') x P(A')
        1. P(A') = 1 – P(A) = 1 - .30 = .70
        2. P(B) = (0.40 x 0.30) + (0.10 x 0.70) = 0.19
    iii. P(A|B) = (P(B|A) x P(A)) / P(B) = (0.40 x 0.30) / 0.19 = 0.6316 = 63.16%
    iv. Given that they have a PhD, the probability that an employee is a Data Scientist is 63.16%.

Problem 4: Law of Total Probability

A diagnostics test has a probability of 0.95 giving a positive result when applied to a person suffering from a certain disease. It has a probability of 0.10 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?

a. A = Person actually has the disease & B = Person tests positive
    i. P(A) = 0.005 (0.5% of the population has this disease)
    ii. P(B|A) = 0.95 (Probability of a positive test result given that the person has the disease)
    iii. P(B|A') = 0.10 (Probability of a positive test result given that the person does not have the disease)
b. Bayes' Theorem: P(A|B) = (P(B|A) x P(A)) / P(B)
c. Use Law of Probability to find P(B): P(B) = P(B|A) x P(A) + P(B|A') x P(A')
    i. P(A') = 1 - P(A) = 1 - 0.005 = 0.995
    ii. P(B) = (0.95 x 0.005) + (0.10 x 0.995) = 0.00475 + 0.0995 = 0.10425
d. P(A|B) = (P(B|A) x P(A)) / P(B) = (0.95 x 0.005) / 0.10425 = 0.00475 / 0.10425 = 0.0455
e. If a person tested positive in the test, the probability that the person actually has the disease is 4.55%.

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

a. Calculated P(B) earlier: P(B) = (0.95 x 0.005) + (0.10 x 0.995) = 0.00475 + 0.0995 = 0.10425 = 10.425%
b. The total probability of a person testing positive is 10.43%.

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?

a. A = Spam & B = Classified as Spam
    i. P(A) = .90
    ii. P(B|A) = 0.95
b. Bayes' Theorem: P(A|B) = (P(B|A) x P(A)) / P(B)
c. Use Law of Probability to find P(B): P(B) = P(B|A) x P(A) + P(B|A') x P(A')
    i. P(A') = 1 - P(A) = 1 - 0.90 = 0.10
    ii. P(B) = (0.95 x 0.90) + (0.15 x 0.10) = 0.855 + 0.015 = 0.870
d. P(A|B) = (P(B|A) x P(A)) / P(B) = (0.95 x 0.90) / 0.870 = 0.855 / 0.870 = 0.9828 = 98.28%
e. If an email is picked at random, and the filter classifies it as spam, the probability that it is actually spam is 98.28%.

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

a. A = Spam & B = Not Classified as Spam
    i. P(A) = .90
    ii. P(B|A') = 0.85
b. Bayes' Theorem: P(A|B) = (P(B|A) x P(A)) / P(B)
c. Use Law of Probability to find P(B): P(B) = P(B|A) x P(A) + P(B|A') x P(A')
    i. P(A') = 1 - P(A) = 1 - 0.90 = 0.10
    ii. P(B) = (0.15 x 0.10) + (0.85 x 0.90) = 0.015 + 0.765 = 0.780
d. P(A|B) = (P(B|A') x P(A')) / P(B) = (0.15 x 0.10) / 0.780 = 0.015 / 0.780 = 0.0192 = 1.92%
e. If an email is classified as "not spam", the probability that it is actually spam is 1.92%.

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.

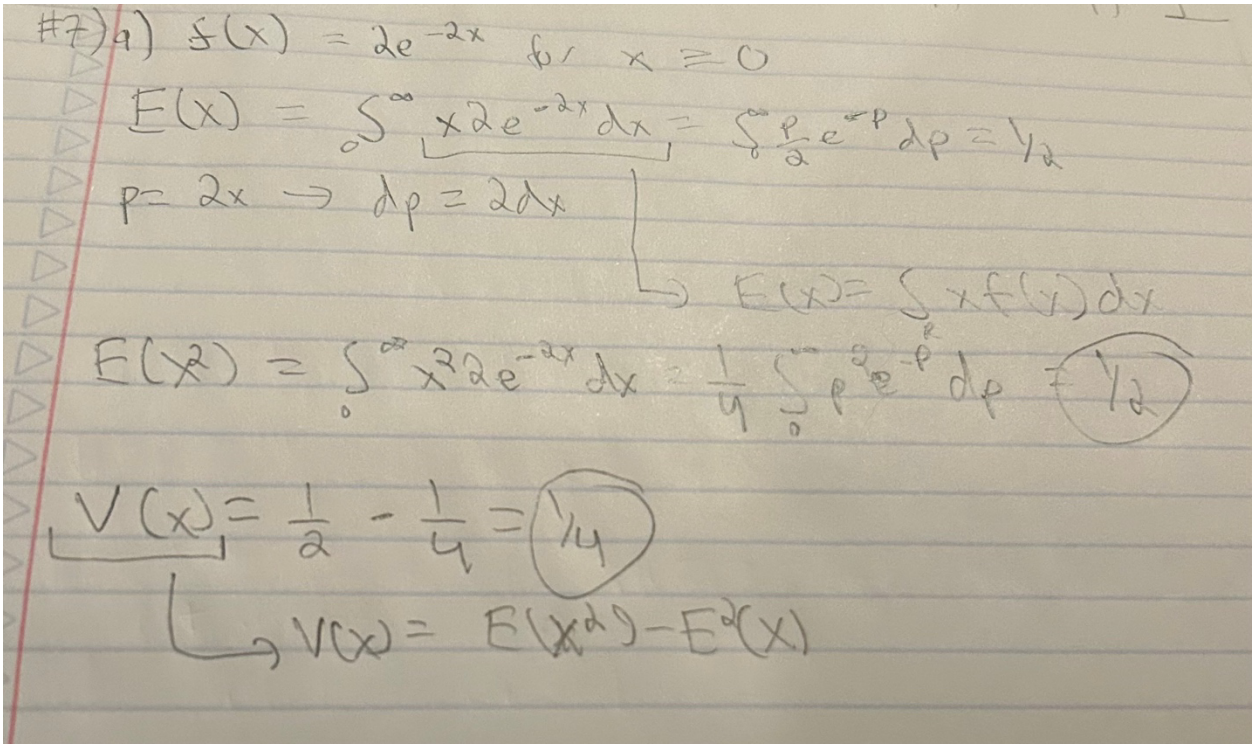
a. X is a random variable that represents the outcome of the game in terms of money won or lost on a single roll of a fair six-sided die.
    i. X = +$10 for rolling a 6
    ii. X = -$2 for any other roll

(b) Compute the expected value of X.

a. P(6) = 1/6 & P(Other #) = 5/6
b. E(X) = (+10) x P(X = +10) + (-2) x P(X = -2)
c. E(X) = (+10) x (1/6) + (-2) x (5/6) = 10/6 - 10/3 = 5/3 - 20/6 = (10 - 20)/6 = -10/6 = -5/3 = -$1.67
d. The expected value of X is -$1.67.

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 >= 0.



(a) Compute the expected value E[X] of X.
   a. E[X] = 1/2

(b) Compute the variance Var[X] of X.
   a. Var[X] = 1/4

(c) Interpret your findings from parts (a) and (b) in the context of the server's processing time.
   a. The average time for the server to process a certain type of query is 30 minutes or a half of an hour.
   b. The variability in the time it takes to process a certain type of query is 15 minutes or a quarter of an hour.

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.

	S	**C**	**R**
S	0.7	0.2	0.1
C	0.3	0.4	0.3
R	0.2	0.3	0.5

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?
   a. P(0) = [1, 0, 0]
   b. P(2) = P(0) x M^2
   c. P(2) = [1, 0, 0] x M^2
      i. Square the transition matrix.
   d. P(2) = [1, 0, 0] x M^2 = [1 x 0.52 + 0 x 0.38 + 0 x 0.35, 1 x 0.29 + 0 x 0.34 + 0 x 0.35, 1 x 0.19 + 0 x 0.28 + 0 x 0.3] = [0.52, 0.29, 0.19]
   e. The probability that it will be rainy 2 days from now is 19%.

(c) Find the stationary distribution of this Markov chain.
   a. Row vector π = [π(S), π(C), π(R)]
      i. [π(S), π(C), π(R)] x M = [π(S), π(C), π(R)]
   b. π(S) = π(S) * 0.7 + π(C) * 0.3 + π(R) * 0.2
      i. π(S) = 0.375
   c. π(C) = π(S) * 0.2 + π(C) * 0.4 + π(R) * 0.3
      i. π(C) = 0.375
   d. π(R) = π(S) * 0.1 + π(C) * 0.3 + π(R) * 0.5
      i. π(R) = 0.25
   e. The stationary distribution of this Markov chain is π = [0.375, 0.375, 0.25].

(d) Interpret the stationary distribution in the context of this weather model.
   a. According to the stationary distribution in the context of this weather model on average, 37.5% of the time the weather will be sunny, 37.5% of the time the weather will be cloudy, and 25% of the time it will be rainy.

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.
   a. Prior distribution: f(θ) = θ^(2-1) x (1-θ)^(2-1) / B(2, 2)
   b. Posterior distribution ∝ Likelihood x Prior distribution
   c. Likelihood: L(θ) = θ^30 x (1-θ)^(100-30)
   d. Posterior distribution ∝ θ^30 x (1-θ)^(100-30) x θ^(2-1) x (1-θ)^(2-1) / B(2, 2) = θ^(30+2-1) x (1-θ)^(100-30+2-1) / B(2, 2)
   e. Posterior distribution: Beta(30+2, 100-30+2) = Beta(32,72)

(b) Based on the posterior distribution, provide an estimate for θ.
   a. Mean of posterior distribution = α / (α + β)
   b. Estimated θ = 32 / (32 + 72) = 32/104 = 0.3077 = 30.77%
   c. Estimated θ (or the recovery rate) = 30.77%.

(c) Explain the role of the conjugate prior in simplifying the calculation of the posterior distribution.
   a. The conjugate prior simplifies the calculation of the posterior distribution in Bayesian inference by ensuring that the posterior distribution belongs to the same family of probability distributions as the prior distribution. In the case of the Beta distribution as a conjugate prior for the Bernoulli distribution, updating the prior distribution using Bayes' theorem results in a posterior distribution that is still a Beta distribution. By directly updating the parameters of the prior distribution with observed data, complex expressions and numerical integration can be avoided, making the calculation process more straightforward. The conjugate prior serves as a mathematically convenient choice for updating beliefs based on data.

Extra Credit: Non-Informative Priors (20 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.

a. ...
b. ...
c. ...
d. Using non-informative priors in Bayesian inference has several pros. One major benefit is that they allow the data to play a more prominent role in the analysis. By minimizing the influence of subjective beliefs or assumptions in the prior distribution, non-informative priors can provide a more impartial approach to statistical inference. They allow the evidence contained in the observed data to have a stronger impact on the resulting posterior distribution, leading more reliable conclusions. Additionally, non-informative priors can simplify the computation of posterior distributions by avoiding complex prior specifications. This simplification can be especially valuable when dealing with complex models or when prior information is scarce or difficult to quantify precisely.
   However, there are some cons when using non-informative priors. One significant drawback is that improper non-informative priors can yield improper posterior distributions. Improper priors do not integrate to a finite value, and their resulting posteriors may not be well-defined or interpretable in a conventional sense. Additionally, non-informative priors may not always be truly non-informative in every context. They might still exert some influence on the posterior distribution, particularly

when the sample size is small or when strong prior knowledge or external information is available.