Ainsley McLaughlin 9/18/24

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? $P(DS \cap PhD)$

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

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 1: Total $= 3+2+5 = 10$

a) $P(\text{red}) = 3/10 = \boxed{0.3}$

b) $P(\text{blue} \mid \text{blue drawn}) = 4/9 = \boxed{0.444}$

Problem 2: $P(\text{server down}) = 0.5$

a) $P(\text{server1} \cap \text{server2}) \overset{\text{indpt}}{=} P(\text{server1}) \cdot P(\text{server2}) = 0.5 \cdot 0.5 = \boxed{0.25}$

b) $P(\geq 1 \text{ server down}) = 1 - P(< 1 \text{ server down}) = 1 - P(\text{both servers are up}) \Rightarrow$

* Note: $P(\text{one server up}) = 1 - 0.5 = 0.5$, and since independent events:

$P(\text{server1 up and server2 up}) \overset{\text{indpt}}{=} 0.5 \cdot 0.5 = 0.25$

$\Rightarrow 1 - 0.25 = \boxed{0.75}$

Problem 3: Have: $P(DS) = 0.3$ $\quad P(DS^c) = 1 - 0.3 = 0.7$

$P(PhD \mid DS) = 0.4$ $\quad P(PhD \mid DS^c) = 0.1$

a) Want: $P(DS \text{ and } PhD) = P(DS \cap PhD) = P(PhD \mid DS) \cdot P(DS)$

$= 0.4 \cdot 0.3 = 0.12$

Principle: $P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$

b) Want $P(DS \mid PhD)$

Principle: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

$P(B) = P(A \cap B) + P(A^c \cap B)$

$= P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)$

$$P(DS \mid PhD) = \frac{P(DS \cap PhD)}{P(PhD)} = \frac{0.12}{0.19} = \boxed{0.6316}$$

$P(PhD) = P(PhD \mid DS) \cdot P(DS) + P(PhD \mid DS^c) \cdot P(DS^c)$

$= 0.4 \cdot 0.3 + 0.1 \cdot 0.7$

$= 0.19$

Problem 4: Given : $P(\oplus | D) = 0.95$ $P(D) = 0.005$

$P(\oplus | D^c) = 0.1$

Know : $P(D^c) = 1 - P(D) = 0.995$

a) Want $P(D | \oplus) = \dfrac{P(\oplus | D) \cdot P(D)}{P(\oplus)} \Rightarrow$

Where: $P(\oplus) = P(\oplus | D) P(D) + P(\oplus / D^c) P(D^c)$

$= 0.95 \cdot 0.005 + 0.1 \cdot 0.995$

$= 0.10425$

$\Rightarrow \dfrac{0.95 \cdot 0.005}{0.10425} = \dfrac{0.00475}{0.10425}$

$\Rightarrow \boxed{0.04556}$

b) $P(\oplus)$ seen circled in pink above

$= \boxed{0.10425}$

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? $P(S|+)$

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

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.

Problem 5: Given: $P(S) = 0.9$ $\quad P(\oplus|S) = 0.95$

$P(\ominus|S^c) = 0.85$

a) $P(S|\oplus) = \dfrac{P(\oplus|S)\,P(S)}{P(\oplus)} = \dfrac{0.95 \cdot 0.9}{P(\oplus)} = \dfrac{0.95 \cdot 0.9}{0.94} = \boxed{0.901}$

Where: $P(\oplus) = P(\oplus|S) \cdot P(S) + P(\oplus|S^c) \cdot P(S^c)$

$= 0.95 \cdot 0.9 + 0.85 \cdot (1 - 0.9) = 0.94$

b) want $P(S|\ominus) = \dfrac{P(\ominus|S)\,P(S)}{P(\ominus)}$

Note: $P(A|B) = 1 - P(A^c|B)$

So: $P(\ominus|S) = 1 - P(\oplus|S) = 1 - 0.95 = 0.05$

lso: $P(\ominus) = P(\ominus|S) \cdot P(S) + P(\ominus|S^c) \cdot P(S^c)$

$= 0.05 \cdot 0.9 + 0.85 \cdot (1 - 0.9)$

$= 0.13$

Finally: $P(S|\ominus) = \dfrac{0.05 \cdot 0.9}{0.13} = \boxed{0.346}$

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. X = \$ you win/loose
(b) Compute the expected value of X.

a)

$$P(X=10) = P(6 \text{ on dice}) = \frac{1}{6}$$

$$P(X=-2) = P(\text{not } 6 \text{ on dice}) = 1-\frac{1}{6} = \frac{5}{6}$$

$$X = \begin{cases} 10 & \text{, with prob } \frac{1}{6} \\ -2 & \text{, with prob } \frac{5}{6} \end{cases}$$

pmf of X = what you win:

$$\left.\begin{array}{l} P_X(10) = \frac{1}{6} \\ P_X(-2) = \frac{5}{6} \end{array}\right\} \qquad P_X(x) = \begin{cases} \frac{1}{6} & , x=10 \\ \frac{5}{6} & , x=-2 \end{cases}$$

b)

$$E(X) = \sum_{x_k \in R_X} x_k \cdot P(X=x_k)$$

$$= 10 \cdot \frac{1}{6} + (-2) \cdot \frac{5}{6} = 0 \, *$$

*So, as you played this game towards an infinite number of times, on average, you would not gain or loose mone. Your net would be 0

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

a) $E(x) = \int_0^\infty x \cdot f(x)\, dx$ where $f(x) = $ pdf

$= \int_0^\infty x\,(2e^{-2x})\, dx = \int_0^\infty 2x e^{-2x}\, dx$

Integration by parts: $\int u\, dv = uv - \int v\, du$

$u = 2x$ $dv = e^{-2x}$

$du = 2\, dx$ $v = -\frac{1}{2} e^{-2x}$

$= 2x\left(-\frac{1}{2}e^{-2x}\right)\Big|_0^\infty - \int_0^\infty \left(-\frac{1}{2}e^{-2x}\right)(2)\, dx$

$= 0 - \int_0^\infty -e^{-2x}\, dx$

$= -\left[\frac{1}{2}e^{-2x}\Big|_0^\infty\right] = -\left[0 - \frac{1}{2}\right] = \frac{1}{2}$

b) $Var(X) = E[X^2] - (E[X])^2$

$(E[X])^2 \to \left(\frac{1}{2}\right)^2 \to \frac{1}{4}$

$$\int_0^\infty x^2 f(x)\,dx = \int_0^\infty x^2 \left(2e^{-2x}\right) dx$$

$$= \int_0^\infty 2x^2 e^{-2x}\, dx$$

$u = 2x^2 \qquad dv = e^{-2x}$

$du = 4x\, dx \qquad v = -\frac{1}{2} e^{-2x}$

$$= \left[2x^2\left(-\frac{1}{2}e^{-2x}\right)\right]\Big|_0^\infty - \int_0^\infty -\frac{1}{2}e^{-2x}(4x)\,dx$$

$$= 0 - \int_0^\infty -2xe^{-2x}\,dx$$

$$= 0 + \underbrace{\int_0^\infty 2xe^{-2x}}_{\text{from above in a)}} = \frac{1}{2}$$

$$Var(x) = \frac{1}{2} - \frac{1}{4} = \boxed{\frac{1}{4}}$$

c) Interpretation:

For expected value: On average, over many many instances, we would expect the time it would take for the server to process the query would be a half hour.

For variance: Since the variance is 0.25, which is relatively low, this means the processing time is fairly consistently around the expected average processing time. We could also use the 68-95-99 rule to say that

(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.

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.

a) Transition matrix:

* tomorrow

$$\begin{array}{cccc} & s & c & r \\ \text{sunny} & 0.7 & 0.2 & 0.1 \\ \text{cloudy} & 0.3 & 0.4 & 0.3 \\ \text{rainy} & 0.2 & 0.3 & 0.5 \end{array} = T$$

c) Trying to find vector $X_0 = [x_s, x_c, x_r]$

So $\vec{x}_0 T = \vec{x}_0$

Set up: $[x_s, x_c, x_r]\begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} = [x_s, x_c, x_r]$

And $x_s + x_c + x_r = 1$ (prob vector)

System:
$$0.7x_s + 0.3x_c + 0.2x_r = x_s$$
$$0.2x_s + 0.4x_c + 0.3x_r = x_c$$
$$0.1x_s + 0.3x_c + 0.5x_r = x_r$$
$$x_s + x_c + x_r = 1$$

Wolfrom-Alpha solution:

$$\vec{x}_0 = \left[\frac{21}{46}, \frac{13}{46}, \frac{12}{46}\right]$$

b) For two days, we first:

$$T \times T = T^2$$

$$\begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix} \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

$$= \begin{pmatrix} 0.57 & 0.25 & 0.18 \\ 0.39 & 0.31 & 0.3 \\ 0.33 & 0.31 & 0.36 \end{pmatrix} = T^2$$

Now only keep what's associated with sunny:

$$T^2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.57 \\ 0.39 \\ 0.33 \end{pmatrix}$$

d) Interpretation of the stationary distribution:

The stationary distribution is essentially like the expected value. It is the proportion of sunny, cloudy, and rainy days you would expect if you tracked the weather over many days. For this case, we would expect that there would be about 21/46 sunny days on average over time, 13/46 cloudy days on average over time and 12/46 rainy days on average over time.

* Sorry this is out of order; I didn't have the energy to rewrite it all.

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:

$$P(x|\theta) = \theta^{x}(1-\theta)^{1-x}, \quad x \in (0,1), \quad 0 \le \theta \le 1$$

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

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

where B(α, β) 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. $\hat{\theta} = 3/10$

(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.

a) Since bernoulli is one trial in binomial, the beta dist is still the conj. prior for θ.
Given prior = Beta(α, β), the posterior is Beta(α+k, β+n−k)
Here k=30, n=100 so
The posterior is Beta(2+30, 2+100−30)
= Beta(32, 72)

b) $E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{32}{32+72} \approx 0.308$

c) The conjugate prior ensures that the posterior distribution will be in the same family.
In this example, since the conjugate prior is Beta, the posterior will also be Beta. It simplifies the process and the interpretation. For example, we can find the closed form expected value as above since the expected value has a known formula based on the parameters of the distribution.

A prior $P(\theta)$ is conjugate if the posterior $P\ \theta\ X$ is of the same distribution.
A prior is proper if it is a well defined probability distribution.
• Advantages:
• Analytically tractable – Calculating posterior probabilities is simple.
• Computationally simple – No greedy estimation, everything is solvable directly
• Very useful in more complex estimation – Simplifies parts of the problem…

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.

a) Since the conjugate prior is $N(0, 100^2)$, the posterior will also be normal.

We want $P(\mu | X, \sigma^2)$ where σ^2 is fixed.

From Bayes:

$$P(\mu | X, \underbrace{\sigma^2}_{\text{fixed}}) = \frac{\overbrace{P(X|\mu)}^{\text{likelihood}} \overbrace{P(\mu)}^{\text{prior}}}{\underbrace{P(X)}_{\text{marginal}}}$$

$$P(\mu \mid x) \propto P(x \mid \mu)\, P(\mu)$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)} \qquad \frac{1}{\sqrt{2\pi(100)^2}} e^{\left(\frac{-(\mu-0)^2}{2\cdot 100^2}\right)}$$

$$\Rightarrow N(0, 100^2)$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)} \cdot \frac{1}{\sqrt{2\pi(100)^2}} e^{\left(\frac{-\mu^2}{2\cdot 100^2}\right)}$$

$$\Rightarrow \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{1}{\sqrt{2\pi(100^2)}}\right) e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) - \frac{\mu^2}{2\cdot 100^2}}$$

$$\Rightarrow \quad // \quad e^{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu\right) - \frac{\mu^2}{2\cdot 100}}$$

$$\Rightarrow \quad // \quad e^{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\mu}{\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2} - \frac{\mu^2}{2\cdot 100}}$$

$$\Rightarrow \quad // \quad e^{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\mu}{\sigma^2}\sum x_i - \frac{n^2}{2}\left(\frac{n}{\sigma^2} + \frac{1}{100}\right)}$$

From powerpoint: posterior:

with $\sigma = 5$ $\quad \mu_0 = 0$

$\sigma_0 = 100$

$$\mu_{post} = \frac{\sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2} + \frac{n(100^2)}{n(100^2)}\bar{X}$$

$$= 0 + \bar{X}$$

$$= \bar{X}$$

$$\sigma^2_{post} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{1}{\frac{n}{25} + \frac{1}{100}}$$

$$= \frac{1}{\frac{4n+1}{100}} = \frac{100}{4n+1}$$

$$\text{Posterior} \sim N\left(\bar{X}, \frac{100}{4n+1}\right)$$

b) Prior: $P(\mu) = 1$, $-\infty < \mu < \infty$

Posterior:

likelihood

$$P(\mu \mid X, \sigma^2) \propto P(X \mid \mu)\, P(\mu)$$

σ^2 known

$P(X \mid \mu)$: same as above

$P(\mu)$: 1

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)}$$

So: $\mu_{post} = \overline{X}$

$$\sigma^2_{post} = \frac{\sigma^2}{n} = \frac{25}{n}$$

So Posterior $\sim N\left(\overline{X}, \frac{25}{n}\right)$

c) These priors differ in that the Jeffey's prior results in a posterior that allows the data to dominate more and depends on the data more since it is initially non-informative. The variance decreases more quickly as n get larger since it does not include any more uncertainty from the prior. For the vague prior, the posterior variance depends on the prior variance and n, making the prior have more influence when n is small. As n gets larger, the influence of the prior decreases, but it converges more slowly than the Jeffrey prior.

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