CS109 March 14, 2024

Section 9: Final Exam Review

Before you leave lab, make sure you click here so that you're marked as having attended this week's section. The CA leading your discussion section can enter the password needed once you've submitted. **Note:** This is your last section this quarter, so be sure to say goodbye to everyone in your section and wish them well on the final and over the summer.

1 Warmups

1.1 Maximum A Posteriori

a. Intuitively, what is MAP? What problem is it trying to solve? How does it differ from MLE?

b. Given a 6-sided die (possibly unfair), you roll the die N times and observe the counts for each of the 6 outcomes as $n_1, ..., n_6$. What is the maximum a posteriori estimate of this distribution, using Laplace smoothing? Recall that the die rolls themselves follow a multinomial distribution.

1.2 Naive Bayes Review

Recall the classification setting: we have data vectors of the form $X = (X_1, \ldots, X_m)$ and we want to predict a label $Y \in \{0, 1\}$.

a. Recall in Naive Bayes, given a data point x, we compute $P(Y = 1|X = x)$ and predict $Y = 1$ provided this quantity is ≥ 0.5, and otherwise we predict $Y = 0$. Decompose $P(Y = 1|X = x)$ into smaller terms, and state where the Naive Bayes assumption is used.

b. Suppose we are given example vectors with labels provided. Give a formula to estimate (using maximum likelihood) each quantity $P(X_i = x_i|Y = y)$ above, for $i \in \{1, \ldots, m\}$ and $y \in \{0, 1\}$. You can assume there is a function count which takes in any number of boolean conditions and returns a count over the data of the number of examples in which they are true. For example, count$(X_3 = 2, X_5 = 7)$ returns the number of examples where $X_3 = 2$ and $X_5 = 7$.

2 Problems

2.1 Bayesian Carbon Dating

We are able to know the age of ancient artefacts using a process called carbon dating. This process involves a lot of uncertainty! Living things have a constant proportion of a molecule called C14 in them. When living things die those molecules start to decay. The time to decay in years, T, of a C14 molecule is distributed as an exponential. $T \sim \text{Exp}(\lambda = 1/8267)$.

a. Consider a single C14 molecule. What is the probability that it decays within 500 years?

b. C14 molecules decay independently. A particular sample started with 100 molecules. What is the probability that exactly 95 are left after 500 years? Let p be your answer to part a.

c. Write pseudocode for a function `pr_measure_given_age(m, age)` which returns $P(M = \texttt{m} | A = \texttt{age})$, the probability that exactly `m` molecules are left out of the original 100 after exactly `age` number of years.

d. You observe a measurement of 95 C14 molecules in a sample. You assume that the sample originally had 100 C14 molecules when it died. Write pseudocode for a function `age_belief()` that returns a list of length 1000 where the value at index i in the list stores $P(A = i | M = 95)$. Age is a discrete random variable which takes on whole numbers of years. $A = i$ is the event that the sample organism died i years ago. You may use the function `pr_measure_given_age(m, age)` from part c. For your prior belief: you know that the sample ***must*** be between $A = 500$ and $A = 600$ inclusive and you assume that every year in that range is equally likely.

2.2 Continuous Joint Distributions

a. Let X, Y, and Z be independent Normal variables with means of $\mu_X = 4$, $\mu_Y = 5$, and $\mu_Z = 6$ and variances $\sigma_X^2 = 16$, $\sigma_Y^2 = 25$, and $\sigma_Z^2 = 36$. Let $A = X + Y$ and $B = Y + Z$. It can be shown that the joint distribution (A, B) is Bivariate Normal. What are the parameters of the joint distribution (A, B)?

b. Suppose hundreds of thousands (that is, a sufficiently large number) of student scores on a 150-question exam are distributed according to the following random variable:

$$R = \sum_{i=1}^{50} M_i + 0.5 \sum_{j=1}^{100} W_j \tag{1}$$

Each of the M_i are independent and identically distributed (IID) Beta random variables—yes, the questions are scored on a continuous scale from 0 to 1—and the W_j are separate IID Beta random variables, where all W_j are independent of all M_i. The Beta parameters are $\alpha_M = 10$, $\beta_M = 2$, $\alpha_W = 8$, and $\beta_W = 4$. If we sample 100 student scores $R_1, \ldots, R_n$ IID according to the distribution of R above, what is the distribution of the sample mean $\overline{R}$?

2.3 Timing Attacks

In this problem we are going to show you how to crack a password in linear time, by measuring how long the password check takes to execute (see code below). Assume that our server takes T ms to execute any line in the code where $T \sim N(\mu = 5, \sigma^2 = 0.5)$ ms. The amount of time taken to execute a line is always independent of other values of T.

```
# An insecure string comparison
def string_equals(guess, password):
    n_guess = len(guess)
    n_password = len(password)
    if n_guess != n_password:
        return False                # 4 lines executed to get here
    for i in range(n_guess):
        if guess[i] != password[i]:
            return False            # 6 + 2i lines executed to get here
    return True                     # 5 + 2n lines executed to get here
```

On our site all passwords are length 5 through 10 (inclusive) and are composed of lower case letters only. A hacker is trying to crack the root password which is "gobayes" by carefully measuring how long we take to tell them that her guesses are incorrect.

a. What is the distribution of time that it takes our server to execute k lines of code? Recall that each line independently takes $T \sim N(\mu = 5, \sigma^2 = 0.5)$ ms.

b. First the hacker needs to find out the length of the password. What is the probability that the time taken to test a guess of correct length (server executes 6 lines) is longer than the time taken to test a guess of an incorrect length (server executes 4 lines)? Assume that the first letter of the guess does not match the first letter of the password. Hint: $P(A > B)$ is the same as $P(A - B > 0)$.

c. Now that our hacker knows the length of the password, to get the actual string she is going to try and figure out each letter one at a time, starting with the first letter. The hacker tries the string "aaaaaaa" and it takes 27ms. Based on this timing, how much more probable is it that first character did not match (server executes 6 lines) than the first character did match (server executes 8 lines)? Assume that all letters in the alphabet are equally likely to be the first letter.

d. If it takes the hacker 6 guesses to find the length of the password, and 26 guesses per letter to crack the password string, how many attempts does she need to crack our password, "gobayes"? Yikes!

2.4 Naïve Bayes

Suppose we observe two discrete input variables X_1 and X_2 and want to predict a single binary output variable Y (which can have values 0 or 1). We know that the functional forms for the input variables are $(X_1|Y = 0) \sim \text{Poi}(\lambda_0)$, $(X_1|Y = 1) \sim \text{Poi}(\lambda_1)$, $(X_2|Y = 0) \sim \text{Ber}(p_0)$, and $(X_2|Y = 1) \sim \text{Ber}(p_1)$, but we don't know the optimal values of the parameters. We are, however, given a dataset of 9 training instances (shown at right.)

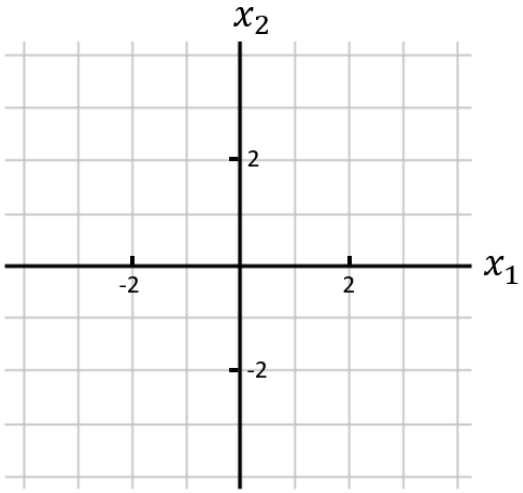
X_1	X_2	Y	X_1	X_2	Y
1	1	**0**	3	1	**1**
3	0	**0**	5	0	**1**
7	1	**0**	5	1	**1**
9	0	**0**	5	1	**1**
			7	1	**1**

a. Use Maximum Likelihood Estimation to estimate the parameters λ_0, p_0, λ_1, and p_1.

b. Use Maximum Likelihood Estimation to estimate the parameter p_y for $Y \sim Ber(p_y)$.

c. You observe the following testing instance: $(X_1, X_2) = (2, 0)$. Using the Naïve Bayes assumption, predict the output Y for the testing instance. For this problem, showing how you computed your prediction is worth more points than the final answer.

2.5 Logistic regression



Suppose you have trained a logistic regression classifier that accepts as input a data point (x_1, x_2) and predicts a class label $\hat{Y}$. The parameters of the model are $(\theta_0, \theta_1, \theta_2) = (2, 2, -1)$. On the axes, draw the decision boundary $\theta^T \mathbf{x} = 0$ and clearly mark which side of the boundary predicts $\hat{Y} = 0$ and which side predicts $\hat{Y} = 1$.

2.6 The Most Important Features

Let's explore saliency, a measure of how important a feature is for classification. We define the saliency of the ith input feature for a given example $(\mathbf{x}, y)$ to be the absolute value of the partial derivative of the log likelihood of the sample prediction, with respect to that input feature $\left|\frac{\partial LL}{\partial x_i}\right|$. In the images below, we show both input images and the corresponding saliency of the input features (in this case, input features are pixels):

First consider a trained logistic regression classifier with weights θ. Like the logistic regression classifier that you wrote in your homework it predicts binary class labels. In this question we allow the values of $\mathbf{x}$ to be real numbers, which doesn't change the algorithm (neither training nor testing).

a. What is the Log Likelihood of a single training example $(\mathbf{x}, y)$ for a logistic regression classifier?

b. Calculate is the saliency of a single feature (x_i) in a training example $(\mathbf{x}, y)$.

c. Show that the ratio of saliency for features i and j is the ratio of the absolute value of their weights $\frac{|\theta_i|}{|\theta_j|}$.

3 Ethics and Beta Distribution

While there won't be any ethics material on the final exam, we're including a problem that will not only exercise some probability, but hopefully provoke you to begin thinking about the impact that probability- and data-driven decisions have on society.

The Economist used a beta distribution to forecast results for the 2020 U.S. presidential election.*

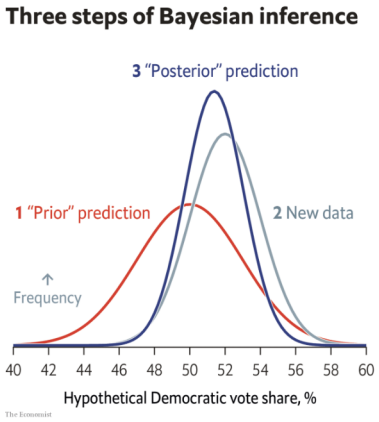


Figure 1: Updated prediction of Democratic vote share is "Posterior" prediction.

*Gelman, A., & Heidemanns, M. (2020). How the economist presidential forecast works. The Economist.

1. Why is the beta distribution appropriate for modeling a presidential election?

2. Read the polling report published by The Economist. What should be considered when using this model and releasing its election predictions?