ICS 635 HW2

February 21, 2019

Bayesian Reasoning

Problem 1: True Detective

Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type O blood. The blood groups of the two traces are found to be of type O (a common type in the local population, having frequency 60%) and of type AB (a rare type, with frequency 1%). Do these data (the blood types found at the scene) give evidence in favor of the proposition that Oliver was one of the two people whose blood was found at the scene? [From McKay, 2003]

Problem 2: A Bent Coin

Suppose we have a bent coin. We treat a single coin flip as a random variable $X \in 0, 1$ where 0,1 correspond to tails, heads respectively. We model X as a sample from a Bernoulli distribution parameterized by $\theta \in [0, 1]$, written as $X \sim$ Bernoulli(θ). Thus a particular value x of random variable X has probability

$$p(x) = \theta^x (1 - \theta)^{(1-x)}$$

Suppose we believe (subjectively) that any value of θ in the interval $[0, 1]$ is equally likely, and we can encode that information using a uniform prior. However, we will formulate the uniform prior as a Beta distribution $\theta \sim Beta(\alpha = 1, \beta = 1)$ for mathematical convenience (the Beta is the conjugate prior of the Bernoulli distribution, and is equivalent to the uniform prior when $\alpha = 1, \beta = 1$). The probability density function (pdf) of the Beta distribution is given by

$$p(\theta) = \frac{1}{Z(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

The *partition* function Z is a normalization factor that makes the pdf integrate to one, I.E. $\int_0^1 d\theta p(\theta) = 1$. For the Beta distribution, this Z function makes use of the Gamma function Γ which generalizes the factorial to real numbers ($\Gamma(n) = (n - 1)!$ when n is a positive integer.).

$$Z(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

You flip the coin once and it comes up heads.

1. What is the maximum likelihood estimate (MLE) of θ?
2. What is the maximum a posterior (MAP) estimate of θ?
3. What is the mean posterior (MP) estimate of θ?
4. What is the posterior of θ, $p(\theta|x)$?
5. What is the probability that the next flip will be heads, $p(x_2 = 1|x = 1, \theta)$?
6. You flip the original coin a total of 5 times and it comes up heads every time. Now what are the MLE, MAP, and MP estimates for θ?
7. Setting the beta prior: Suppose we obtain a new bent coin, and an examination leads us to believe that $E[\theta] = m$ and $var[\theta] = v$. What values of α and β will give our prior these properties? [From Murphy 2012]

Problem 3: Naive Bayes by Hand

Consider a Naive Bayes model (multivariate Bernoulli version) for spam classification with the vocabulary $V = \{secret, offer, low, price, valued, customer, today, dollar, million, sports, is, for, play, healthy, pizza\}$. We have the following example spam messages, "*million dollar offer*", "*secret offer today*", "*secret is secret*" and normal messages, "*low price for valued customer*", "*play secret sports today*", "*sports is healthy*", "*low price pizza*". Give the MLEs for the following parameters: θ_{spam}, $\theta_{secret|spam}$, $\theta_{secret|non-spam}$, $\theta_{sports|non-spam}$, $\theta_{dollar|spam}$. [From Daphne Koller]

Supervised Learning

Problem 4: Train vs. Test Error

1. The error on the test will always *decrease* as we get more training data, since the model will be better estimated. However, for complex models, the error on the training set can *increase* as we get more training data, until we reach some plateau. Explain why.
2. Suppose we have a randomly labeled dataset (i.e., the features x tell us nothing about the class labels y) with N_1 examples of class 1, and N_2 examples of class 2, where $N_1 = N_2$. What is the best misclassification rate any method can achieve? What is the estimated misclassification rate of the same method using Leave-One-Out-Cross-Validation? [From Witten05, p152.]

Problem 5: Quasars

Classify quasars from galaxies and stars, using real data from the Sloan Digital Sky Survey. The following Colab notebook provides example code for a 1-NearestNeighbor classifier. Modify the code to improve the test set classification accuracy using *each* of the following classifiers: KNN, Gaussian Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Decision Trees. Try to get good results for each by tuning the hyperparameters. You don't need to turn in code, but describe for each model (1) the data preprocessing and hyperparameters used and how you chose them, and (2) the expected generalization error, providing enough details to reproduce your results.

Note: Over-estimating generalization performance will be penalized, since this is a common mistake in machine learning.

`https://github.com/peterjsadowski/sklearn_examples/blob/master/sdss/quasars.ipynb`

Solutions

Problem 1

Let D be the data, and S be the event that Oliver was present at the scene.

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)} \quad \text{(Bayes rule)}$$

When comparing two hypotheses (S vs. not S, or $\neg S$), we can write the ratio

$$\frac{P(S|D)}{P(\neg S|D)} = \frac{P(D|S)P(S)}{P(D|\neg S)P(\neg S)}$$

We have no information about the prior $P(S)$; the question only asks whether this evidence supports the hypothesis S, so the question is whether $\frac{P(D|S)}{P(D|\neg S)} > 1$. We can calculate each in turn, but it is easy to make a mistake here; we have to be precise about our logical propositions.

There are two blood samples: X_1, X_2. Let K_1, K_2 represent the identities of the potential killers who correspond to the samples, respectively. Then the data we have observed is that either $(X_1 = O, X_2 = AB)$ or $(X_1 = AB, X_2 = O)$, and proposition $S =$ Oliver $\in \{K_1, K_2\}$.

$$\begin{aligned}
P(D|S) &= P((X_1 = O, X_2 = AB) \text{ or } (X_1 = AB, X_2 = O)|S) \quad \text{(by def. of } D\text{)} \\
&= P(X_1 = O, X_2 = AB|S) + P(X_1 = AB, X_2 = O|S) \quad \text{(prob of distinct events)} \\
&= 2 * P(X_1 = O, X_2 = AB|S) \quad \text{(by symmetry)} \\
&= 2 * P(X_1 = O, X_2 = AB|K_1 = \text{Oliver})P(K_1 = \text{Oliver}|S) \\
&\quad + 2 * P(X_1 = O, X_2 = AB|K_2 = \text{Oliver})P(K_2 = \text{Oliver}|S) \\
&= P(X_1 = O|K_1 = \text{Oliver})P(X_2 = AB|K_1 = \text{Oliver}) \\
&\quad + P(X_1 = O|K_2 = \text{Oliver})P(X_2 = AB|K_2 = \text{Oliver}) \\
&= (1)(0.01) + (0.6)(0) \\
&= 0.01
\end{aligned}$$

Now we can split the first term as follows. The second term is equivalent by symmetry:

$$\begin{aligned}
P(D|\neg S) &= P(X_1 = O, X_2 = AB|\neg S) + P(X_1 = AB, X_2 = O|\neg S) \\
&= (0.6)(0.01) + (0.01)(0.6) \\
&= 0.012
\end{aligned}$$

Thus we can quantify the phlebotomological evidence as $P(D|S)/P(D|\neg S) = 0.01/0.012 \approx 0.83$, suggesting that Oliver is *less* likely to be one of the people at the crime.

Problem 2:

1. $\hat{\theta}_{MLE} \triangleq \underset{\theta}{\operatorname{argmax}}\, p(D|\theta) = 1$ where D is the data representing $x_1 = 1$.

2. To compute $\hat{\theta}_{MAP} \triangleq \underset{\theta}{\operatorname{argmax}}\, p(\theta|D)$, we apply Bayes rule to compute the posterior:

$$\begin{aligned}
p(\theta|D) &= p(D|\theta)p(\theta)/p(D) \\
&= (\theta)(1)/p(D) \\
&= \theta/p(D) \\
&= 2\theta
\end{aligned}$$

Since

$$\int_{\theta=0}^{1} p(D|\theta)p(\theta)d\theta = 0.5$$

Note that this can also be computed a simpler way, by observing that $p(D|\theta)p(\theta)$ has the form of a Beta distribution with $\alpha = 2, \beta = 1$. Then, we know that the partition function is the Beta function $B(2,1) = \frac{\Gamma(2)\Gamma(1)}{\Gamma(2+1)} = \frac{1!+1!}{2!} = \frac{1}{2}$.

The max of this function occurs at $\hat{\theta}_{MAP} = 1$.

3. From above, $p(\theta|D) = 2\theta$. Thus,

$$\begin{aligned}
\hat{\theta}_{MP} &\triangleq E_\theta[p(\theta|D)] \\
&= \int_{\theta=0}^{1} \theta p(\theta|D) d\theta \\
&= \frac{2}{3}\theta^3|_0^1 \\
&= \frac{2}{3}
\end{aligned}$$

4. The posterior is just the function $p(\theta|D) = 2\theta$. Note that this function contains more information than any of the point estimates.

5. The computation here happens to be identical to the MP estimate:

$$\begin{aligned}
p(x_2 = 1|D) &= \int_{\theta=0}^{1} p(x_2 = 1|\theta)p(\theta|D)d\theta \\
&= \int_{\theta=0}^{1} p(x_2 = 1|\theta)p(\theta|D)d\theta \\
&= \int_{\theta=0}^{1} (\theta)(2\theta)d\theta \\
&= \frac{2}{3}\theta^3|_0^1 \\
&= \frac{2}{3}
\end{aligned}$$

6. You flip the original coin a total of 5 times and it comes up heads every time. Now what are the MLE, MAP, and MP estimates for θ?

   The MLE $\hat{\theta}_{MLE} = 1$ again. To compute the posterior $p(\theta|D)$, it is simplest to observe that the posterior is also Beta distribution with $\alpha = 1 + 5, \beta = 1$. Thus, $\hat{\theta}_{MAP} = 1$ and $\hat{\theta}_{MP} = E[Beta(6,1)] = \frac{6}{6+1} = \frac{6}{7}$ by the properties of the Beta distribution.

7. If $\theta \sim Beta(\alpha, \beta)$, then

$$
\begin{aligned}
E[\theta] &= \frac{\alpha}{\alpha+\beta} \\
var[\theta] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

All we have to do is solve for α, β in terms of m, v. We can start by observing that

$$
\begin{aligned}
\alpha+\beta &= \frac{\alpha}{m} \\
\beta &= \alpha(\frac{1}{m}-1)
\end{aligned}
$$

and then

$$
\begin{aligned}
v &= \frac{\alpha(\alpha(\frac{1}{m}-1))}{(\frac{\alpha}{m})^2(\frac{\alpha}{m}+1)} \\
&= \frac{\alpha^2(\frac{1}{m}-1)}{(\frac{\alpha}{m})^2(\frac{\alpha}{m}+1)} \\
&= \frac{\frac{1}{m}-1}{\frac{1}{m^2}(\frac{\alpha}{m}+1)} \\
v(\frac{\alpha}{m}+1) &= m-m^2 \\
\frac{\alpha}{m} &= \frac{m-m^2}{v}-1 \\
\alpha &= \frac{m^2-m^3}{v}-m
\end{aligned}
$$

and thus

$$
\beta = \left(\frac{m^2-m^3}{v}-m\right)(1/m-1).
$$

Problem 3:

Each $\hat{\theta}_{MLE}$ can be estimated using empirical counts.

$$\hat{\theta}_{spam} = \frac{N_{spam}}{N_{spam} + N_{non-spam}} = \frac{3}{7}$$

$$\hat{\theta}_{secret|spam} = \frac{N_{spam\text{ and has "secret"}}}{N_{spam}} = \frac{2}{3}$$

$$\hat{\theta}_{secret|non-spam} = \frac{1}{4}$$

$$\hat{\theta}_{sport|non-spam} = \frac{2}{4} = \frac{1}{2}$$

$$\hat{\theta}_{dollar|spam} = \frac{1}{3}$$

Problem 4:

1. A complex model can achieve very low training error by overfitting; if the training data grows while the complexity of model remains constant, the training error will increase.

2. The best that can be achieved is $\frac{1}{2}$. Using LOOCV, the best we would expect to achieve is $\frac{N_1-1}{2N_1-1}$, because the training data set will be slightly unbalanced, and in the ”wrong” direction.

Problem 5:

This data set is relatively easy to work with, and there is enough data that overfitting isn’t as big of a problem as in HW1. The more powerful methods, such as Random Forest classifiers will be able to get $\approx 99\%$ performance on test set. Performance of the other classifiers will vary depending on hyperparameter choices and data preprocessing.