Problem Set #1
CS F320 – Foundations of Data Science
I Semester 2019-20

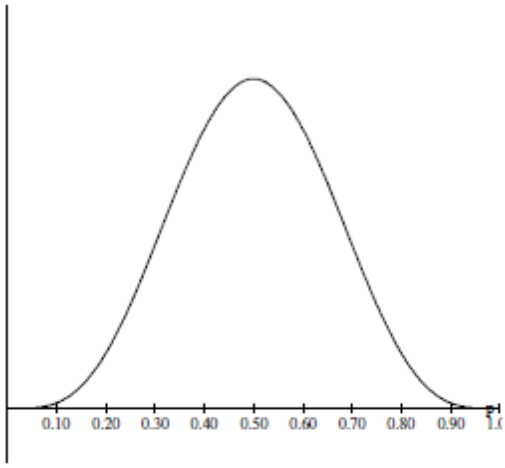
1. The following table consists of training data from an employee database. The data have been generalized. For a given row entry, *count* represents the number of data tuples having the values for *department, status, age,* and *salary* given in that row.

Department	Status	Age	Salary	Count
Sales	Senior	31-35	46-50K	30
Sales	Junior	26-30	26-30K	40
Sales	Junior	31-35	31-35K	40
Systems	Junior	21-25	46-50K	20
Systems	Senior	31-35	66-70K	5
Systems	Junior	26-30	46-50K	3
Systems	Senior	41-45	66-70K	3
Marketing	Senior	36-40	46-50K	10
Marketing	Junior	31-35	41-45K	4
Secretary	Senior	46-50	36-40K	4
Secretary	Junior	26-30	26-30K	6

(a) Given a data sample with values *"systems", "junior",* and *"26-30"* for the attributes *department, status,* and *age*, respectively, what would a Naïve Bayesian Classification of the *salary* for the sample be?
(b) Generate a test set having 30 records. Find the accuracy of NBC.
(c) Apply PCA to the above data to keep 2 most informative attributes.
(d) Use Naïve Bayes' after applying PCA and check if the accuracy has improved.

2. What happens to the volume of a sphere when the number of dimensions grow from 2,3, to say 100? Argue Mathematically.
3. What is the difference between "Likelihood" and "Posterior"? Give some examples to illustrate.
4. I take a coin out of my pocket and I want to estimate the probability of heads when it is tossed. I am only able to toss it 10 times. When I do that, I get seven heads. I ask three statisticians to help me decide on an estimator of *p*, the probability of heads for that coin.
   Case 1. Sue, a frequentist statistician, used p = X/10 = 7/10=0.7
   Case 2. Jose, who doesn't feel comfortable with this estimator, says that he already has an idea that *p* is close to 0.5, but he also wants to use the data to help estimate it. How can he blend his prior ideas and the data to get an estimate?
   Jose makes a sketch of his prior belief about *p*. He thinks it is very unlikely that p is 0 or 1, and quite likely that it is somewhere pretty close to 0.5. He graphs his belief.

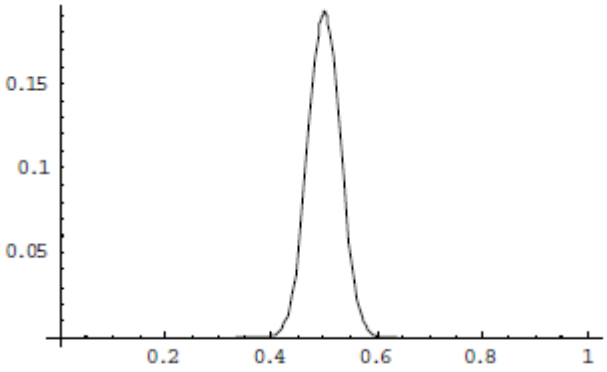
Jose's drawing:



Then he notices that the graph corresponds to a particular probability distribution, which is β (5,5). So this is called his prior distribution for p. (read about β-distribution)

- Suggest how Jose can combine his prior with data to get a distribution for *p?* Show that the posterior is β (12,8).

Case 3: Vicki, who is very sure that coins are unbiased, has a prior distribution like Jose's, but much narrower. There's a much higher probability on values very close to 0.5. She graphs her belief.
Vicki's prior distribution:
She notices that this corresponds to a particular probability distribution, which is Beta(138,138), so that is her prior distribution of p. So the mean is 0.5, the variance is 0.0009, and the standard deviation is 0.03. Notice that her standard deviation is much smaller than Jose's.



- Show that the posterior distribution on Vicki is β (145,141).

Jose and Vicki are both doing Bayesian estimation. Both of them decide to use the mean of the posterior distribution of the parameter as their estimator.

Summary:

- Sue's estimate of the probability of heads: 0.700
- Jose's estimate of the probability of heads: 0.600
- Vicki's estimate of the probability of heads: 0.507

Now, Jennifer offers you a bet. You pick one of these values. She chooses one of the other two. An impartial person tosses the coin 1000 times, and get a sample proportion of heads. If that sample proportion is closer to Jennifer's value, you pay her INR 2000. If it is closer to yours, she pays you INR 2000.

- Which value would you choose and why?

- What are the criticisms of the Bayesian Approach and how they can be addressed?

- Find out what are conjugate priors?

- Find out what are non-informative or flat priors? What is objective Bayesian analysis?

How would your prior belief compare with Jose's or Vicki's? Decide on a mean and standard deviation that reflect your beliefs. Then compute which beta distribution reflects those beliefs. Then compute your posterior distribution, based on your prior and the data. What is the mean of your posterior distribution? That is your Bayesian estimate for *p*, using a conjugate prior distribution that best fits your beliefs. Would you rather make a bet on that, or on one of the three estimates given here?

- What is robust Bayesian analysis?