Naive Bayes

Owen Ward

December 6, 2019

Introduction

- Have talked briefly before about classification problems.
- Have some labeled training data which has some labels. Want to learn a classifier with which we can predict labels for new unlabeled data.
- Even better is a probabilistic classifier

Probabilistic Classifier

- Suppose we have some features X and a label we want to learn Y.
- A probabilistic classifier will give us

$$P(Y|X),$$

  the probability Y will take a certain value, given the features X.
- Classical example is spam emails. X describes the text in the data and a classifier tries to learn the probability an email is a spam email, given the text it contains.
- We will assume Y is binary, so spam email or not, etc.

Bayes Rule

- $P(Y|X)$ is a conditional probability. In these sorts of problems, it gives us a way to relate the quantity we need to things that are easier to compute.
- We can express this rule as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- The denominator ensures this sums to 1 and is a probability.

Bayes Rule

- If A can only take on two values (A, A^c) then

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

- Here $P(A)$ is our prior probability of event A, before we begin observing data.

Bayes Rule

- This rule is extremely useful for calculating complicated conditional probabilities.
- A common example is disease testing.
- Suppose we have a medical test to determine if you have a certain type of cancer. 0.5% of the population have this cancer. If you have cancer, it will correctly determine you have cancer 99% of the time. If you do not have cancer, there is a 2% chance the test will report that you do have cancer.
- Given that you take the test and get a positive result, what is the probability you have cancer?
- Can answer this using Bayes rule.

Disease Testing

- Want to compute

$$P(Disease|PosTest) = \frac{P(PosTest|Disease)P(Disease)}{P(PosTest)}$$

- We know the numbers in the numerator and can compute the denominator.
- $P(PosTest|Disease) = 0.99$
- $P(Disease) = 0.005$

$$P(PosTest) = P(PosTest|Disease)P(Disease)$$

$$+P(PosTest|NoDisease)P(NoDisease)$$

$$= 0.99(0.005) + (0.02)(0.995) = 0.02485.$$

- This gives

$$P(Disease|PosTest) = \frac{0.00495}{0.02485} = 0.199.$$

Naive Bayes

- To use Naive Bayes, we use this method along with another approximation.
- We have
$$P(Y|X) \propto P(X|Y)P(Y)$$
- In applications X is complicated. For spam, it will be the probability all words appear in an email, if it is an spam email. For n words
$$P(X|Y) = P(X_1, X_2, \ldots, X_n|Y).$$
- This can still be very complicated so we make a (naive) assumption that
$$P(X|Y) = P(X_1|Y)P(X_2|Y) \ldots P(X_n|Y),$$
so each feature is actually independent.

Naive Bayes

- So, when we have a binary classifier, we want to estimate the more probable class.
- If the two classes are $Y = 0$ and $Y = 1$ we can compute both

$$P(X_1, X_2, \ldots, X_n | Y = 0)P(Y = 0)$$

  and

$$P(X_1, X_2, \ldots, X_n | Y = 1)P(Y = 1).$$

- Whichever gives the greater value is the more probable class.

Naive Bayes for Text

- To implement this for documents, are $P(X|Y)$ will be the word counts within each document.
- We will actually assume a multinomial model for the word counts. Giving the training data, we calculate the frequency of each word in each class.
- This gives

$$P(X_1, X_2, \ldots, X_n | Y = 1) \propto p_{11}^{X_1} p_{21}^{X_2} \ldots p_{n1}^{X_n},$$

where p is the proportion from the training data and X_i is the number of times word i appeared in the test document.

Naive Bayes for Text

- The previous number is the product of lots of very small probabilities, so we normally do this on the log scale.
- To classify a document, we can just compare

$$\log P(Y) + \log P(X_1, X_2, \ldots, X_n | Y)$$

  for $Y = 0$ and $Y = 1$.
- If we don't know any better will just assume each outcome equally likely initially, so $P(Y = 0) = P(Y = 1) = 0.5$.