10-701/15-781, Machine Learning: Homework 1

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on January 21, 2010

1 Naive Bayes Classifier [30 pt, Field Cady]

In classification problems, we often have several pieces of "evidence". The problem is that, even if each piece of evidence is highly informative, taken together they may be redundant.

Imagine a problem where we must classify data points as y_1 or y_2, based on evidence random variables $X_1, X_2, ..., X_d$. From Bayes rule, we know that $P(y_i|X_1, ..., X_d) \propto P(y_i)P(X_1, ..., X_d|y_i)$, so for classification we just need to compare $P(y_1)P(X_1, ..., X_d|y_1)$ and $P(y_2)P(X_1, ..., X_d|y_2)$.

But calculating $P(X_1, ..., X_d|y_i)$ requires knowing potentially complicated dependencis between the X_i. In a Naive Bayes classifier, we simplify the problem by assuming $P(X_1, ..., X_d|Y) = P(X_1|Y) \times ... \times P(X_d|Y)$; all the pieces of evidence are conditionally independent. And $P(X_i|Y)$ is easy to estimate; if we use the MLE it's just $\frac{\#(X_i \& Y)}{\#Y}$. This classifier is "naive" because somebody might implement it without realizing there could be complicated dependencies, but generally people are aware of its limitations and use it for its simplicity.

1.1 Why We Use Naive Bayes[15 pt]

A big reason we use Naive Bayes classifiers is that they require less training data than Full Bayes Classifiers. This problem should give you a "feel" for how great the disparity really is.

Imagine that each observation is an independent instance of the multi-variate random variable $\vec{X} = X_1, ..., X_d$, where the X_i are i.i.d and Bernoulli(.5). To train a Full Bayes classifier, we need to see every value of $\vec{X}$ "enough" times; training a Naive Bayes classifier only requires seeing both values of X_i "enough" times. We wonder how many observations are needed until, with probability $1 - \epsilon$, we have seen every variable we need to see at least once. To train the classifier well would require more than this, but for this problem we only require one observation.

Hint: You may want to use the following facts:

- For any $k \geq 1$, $(1 - 1/k)^k \leq e^{-1}$
- For *any* events $E_1, ..., E_k$, $Pr\{E_1, ..., E_k\} \leq \sum_{i=1}^{k} Pr\{E_i\}$.

These facts are used a lot when doing probability proofs. The second one, in particular, is called a "union bound" , and we'll use it more later in this course.

1. We start with Full Bayes. Let $\vec{x}$ be a particular value of $\vec{X}$. Show that after N observations, the probability we have never seen $\vec{x}$ is $\leq e^{-N/2^d}$.
2. Using union bounds, show that if more than $N_{FB} = 2^d \ln\left(\frac{2^d}{\epsilon}\right)$ observations have been made, then the probability that *any* value of $\overrightarrow{X}$ has not been seen is $\leq \epsilon$.
3. Now on to Naive Bayes. Show that if N observations have been made, the probability that a given X^i has *not* been seen as both 0 and 1 is $\leq \frac{1}{2^{N-1}}$.
4. Show that if more than $N_{NB} = 1 + \log_2\left(\frac{\epsilon}{d}\right)$ observations have been made, then the probability that *any* X^i has not been observed in both states is $\leq \epsilon$.

5. Let $d = 2$ and $\epsilon = .1$. What are the values of N_{FB} and N_{NB}? What about $d = 5$? And $d = 10$?

1.2 Solution

1. There are 2^d possible observations, all equally likely, so the probability that a particular instance $\overrightarrow{x}$ is not seen in a given trial is $(1 - /frac12^d)$. Since trials are independent, the probability that it hasn't been seen after N trials is $(1 - /frac12^d)^N = \left((1 - /frac12^d)^{2^d}\right)^{N/2^d} \leq (e^{-1})^{N/2^d} = e^{-N/2^d}$

2.
$$\begin{aligned} P(\text{any } \overrightarrow{x} \text{ not seen}) &\leq \Sigma_{\overrightarrow{x}} P(\overrightarrow{x} \text{ not seen}) \\ &= 2^d P(\overrightarrow{x} \text{ not seen}) \\ &\leq 2^d e^{-N_{FB}/2^d} \\ &= 2^d e^{-2^d \ln\left(\frac{2^d}{\epsilon}/2^d\right)} \\ &= 2^d e^{-\ln(2^d/\epsilon)} \\ &= 2^d(\epsilon/2^d) \\ &= \epsilon \end{aligned}$$

3. The only ways it could not be seen are if all observations have had the i^{th} component be 0, or they have all had it be 1. The probabilities of these are each $(1/2)^N$, so the probability that it has not been seen in both states is exactly $(1/2)^N + (1/2)^N = 2(1/2)^N = \frac{1}{2^{N-1}}$.

4. Again using union bounds,
$$\begin{aligned} P(\text{any component not seen in both states}) &\leq \Sigma_i P(\text{component i not seen in both states}) \\ &= dP(\text{a given component not seen in both states}) \\ &= d\frac{1}{2^{N_{NB}-1}} \\ &= d(1/2)^{N_{NB}-1} \\ &= d(1/2)^{\log_2\left(\frac{\epsilon}{d}\right)} \\ &= d1/(\frac{\epsilon}{d}) \\ &= \epsilon \end{aligned}$$

5. If $d = 2$ and $\epsilon = 0.1$ then $N_{FB} = 5.99$ and $N_{NB} = 4.32$. If $d = 5$ and $\epsilon = 0.1$ then $N_{FB} = 184$ and $N_{NB} = 6.6$. If $d = 10$ and $\epsilon = 0.1$ then $N_{FB} = 9455$ and $N_{NB} = 7.6$.

1.3 How bad is Naive Bayes?[15 pt]

Clearly Naive Bayes makes what, in many cases, are overly strong assumptions. But even if those assumptions aren't true, is it possible that Naive Bayes is still pretty good? This problem uses a simple example to explore the limitations of Naive Bayes.

Let X_1 and X_2 be i.i.d. Bernoulli(.5) random variables, and let $Y \in \{1, 2\}$ be some deterministic function of the X^i ; hence Y can be represented by a 2x2 grid of 1s and 2s. Imagine that we have trained a Naive Bayes classifier perfectly, so that $P(Y|X_i)$ is known perfectly.

1. Find a function Y for which the Naive Bayes classifier has a 50% error rate. Given the value of Y, how are X_1 and X_2 correlated?

   We choose

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$Y = 1$	$Y = 2$
$X_1 = 1$	$Y = 2$	$Y = 1$

   To calculate the error rate, we imagine that the observation is (say) $X_0 = 0, X_1 = 0$. Then we compare $P(Y = 1)P(X_0 = 0|Y = 1)P(X_1 = 0|Y = 1) = .5 * .5 * .5$ to $P(Y = 2)P(X_0 = 0|Y = 2)P(X_1 = 0|Y = 2) = .5 * .5 * .5$. They are equal, and we will find the same situation for every other observation. So the Naive Bayes classifier will always say that the values of Y are equally likely, and it must either allways choose a single value (which will make it wrong half the time) or flip a coin (which will also make it wrong half the time).

   Conditioned on knowing the value of Y, either of the X_i uniquely determines the other; this is as strong a violation of the conditional independence assumption as is possible.

2. There are $2^4 = 16$ ways to put value of Y in the grid, but up to symmetry only 4 are unique. Besides the one above there are

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$Y = 1$	$Y = 1$
$X_1 = 1$	$Y = 1$	$Y = 1$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$Y = 1$	$Y = 1$
$X_1 = 1$	$Y = 1$	$Y = 2$

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$Y = 1$	$Y = 2$
$X_1 = 1$	$Y = 1$	$Y = 2$

   In each case, following the procedure above we find that the error rate is 0%.