The Naive Bayes Classifier: Training

Bookmark this page

THE NAIVE BAYES CLASSIFIER: TRAINING (PREFACE)

You should try to answer the following *before* watching the video below which presents the solution.

Note: “log" means natural log in these videos/notes.

Practice problem: Show that the log likelihood for our email spam detection setup can be written as

$$\log\left(\prod_{i=1}^{n} p_{C,Y_1,\ldots,Y_J}(c^{(i)}, y_1^{(i)}, \ldots, y_J^{(i)}; \theta)\right) = f(s) + \sum_{j=1}^{J} g_j(p_j) + \sum_{j=1}^{J} h_j(q_j)$$

for some functions $f, g_1, g_2, \ldots, g_J, h_1, h_2, \ldots, h_J$. What are these functions? Note that what the above equation is saying is that the log likelihood decouples into functions where each function depends on just one of the parameters. Thus, when we want to maximize over θ, what we can do is maximize over each parameter in θ separately! For example, to find what the ML estimate for s is, we only need to look at $f(s)$.

Hint: To show the above equation, you may find it helpful that we can write:

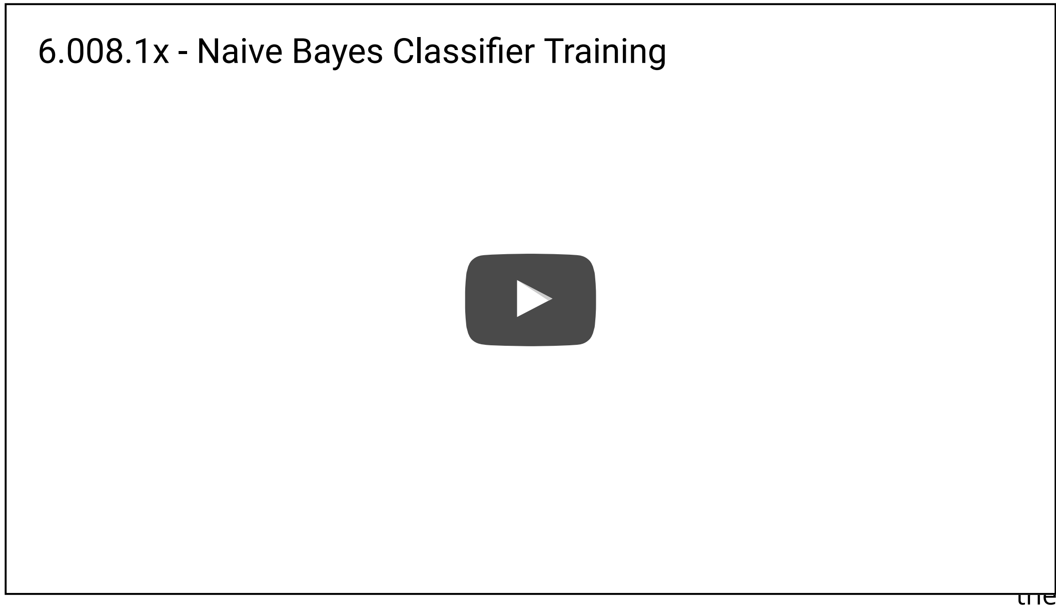
$$p_C(c;\theta) = s^{\mathbf{1}\{c=\text{spam}\}}(1-s)^{1-\mathbf{1}\{c=\text{spam}\}} \qquad \text{for } c \in \{\text{spam}, \text{ham}\}$$
$$p_{Y_j|C}(y_j \mid \text{ham};\theta) = p_j^{y_j}(1-p_j)^{1-y_j} \qquad \text{for } y_j \in \{0,1\}$$
$$p_{Y_j|C}(y_j \mid \text{spam};\theta) = q_j^{y_j}(1-q_j)^{1-y_j} \qquad \text{for } y_j \in \{0,1\}$$

Practice problem: After you figure out what the functions f, g_j's, and h_j's are, obtain the ML estimate for each of the parameters $s, p_1, \ldots, p_J, q_1, \ldots, q_J$ by setting derivatives equal to 0.

Hint: You may find it helpful that for nonzero constants A and B,

$$\frac{d}{dt}\{A\log t + B\log(1-t)\} = 0 \qquad \text{when} \qquad t = \frac{A}{A+B}.$$

The Naive Bayes Classifier: Training



▶ 14:04 / 14:04 1.0x

we're ready to do maximum
lihood estimation.

d in particular, as a reminder,
en we have--

this we saw when I talked
out maximum likelihood

imation before for the biased
n.

whenever we have A log s plus-

A is just some constant that
esn't depend on s--

s B log 1 minus s,

then the derivative of this thing

is equal to 0, when s is equal to A over A plus B.

And so now you can say what is the ML estimate for parameter s?

Well, it's just going to be: here's our A, here's our B

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

These notes cover roughly the same content as the video:

THE NAIVE BAYES CLASSIFIER: TRAINING (COURSE NOTES)

Simplifying the log likelihood: The log likelihood is given by

$$\begin{aligned}
&\log\left(\prod_{i=1}^{n} p_{C,Y_1,\ldots,Y_J}(c^{(i)}, y_1^{(i)}, \ldots, y_J^{(i)}; \theta)\right) \\
&= \log\left(\prod_{i=1}^{n}\left[p_C(c^{(i)};\theta)\prod_{j=1}^{J} p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta)\right]\right) \\
&= \sum_{i=1}^{n}\left[\log p_C(c^{(i)};\theta) + \sum_{j=1}^{J}\log p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta)\right] \\
&= \underbrace{\sum_{i=1}^{n}\log p_C(c^{(i)};\theta)}_{(*)} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{J}\log p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta).}_{(**)}
\end{aligned}$$

We next simplify the expressions (*) and (**).

First let's simplify term (*):

$$\begin{aligned}
(*) = &\sum_{i=1}^{n}\log p_C(c^{(i)};\theta) \\
= &\sum_{i=1}^{n}\left[\mathbf{1}\{c = \text{``spam''}\}\log s + \mathbf{1}\{c = \text{``ham''}\}\log(1-s)\right] \\
= &\left[\sum_{i=1}^{n}\mathbf{1}\{c = \text{``spam''}\}\right]\log s + \left[\sum_{i=1}^{n}\mathbf{1}\{c = \text{``ham''}\}\right]\log(1-s) \\
\triangleq\ & f(s).
\end{aligned}$$

Next, we simplify (**), splitting it up as to decouple p_j and q_j. To do this, we can split the summation over i into two sums, one accounting for all the ham emails and one accounting for all the spam emails:

$$
\begin{aligned}
&(**)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{J}\log p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta)\\
&= \sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``ham''}\}\sum_{j=1}^{J}\log p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta)\\
&\quad + \sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``spam''}\}\sum_{j=1}^{J}\log p_{Y_j|C}(y_j^{(i)}|c^{(i)};\theta)\\
&= \sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``ham''}\}\sum_{j=1}^{J}\left[y_j^{(i)}\log p_j + (1-y_j^{(i)})\log(1-p_j)\right]\\
&\quad + \sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``spam''}\}\sum_{j=1}^{J}\left[y_j^{(i)}\log q_j + (1-y_j^{(i)})\log(1-q_j)\right]\\
&= \sum_{j=1}^{J}\underbrace{\sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``ham''}\}\left[y_j^{(i)}\log p_j + (1-y_j^{(i)})\log(1-p_j)\right]}_{\triangleq g_j(p_j)}\\
&\quad + \sum_{j=1}^{J}\underbrace{\sum_{i=1}^{n}\mathbf{1}\{c^{(i)} = \text{``spam''}\}\left[y_j^{(i)}\log q_j + (1-y_j^{(i)})\log(1-q_j)\right]}_{\triangleq h_j(q_j)}.
\end{aligned}
$$

In summary:

$$f(s) = \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\}\right] \log s + \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“ham ”}\}\right] \log(1-s),$$

$$g_j(p_j) = \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“ham ”}\} y_j^{(i)}\right] \log p_j + \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“ham ”}\}(1 - y_j^{(i)})\right] \log(1-p_j),$$

$$h_j(q_j) = \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\} y_j^{(i)}\right] \log q_j + \left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\}(1 - y_j^{(i)})\right] \log(1-q_j).$$

Setting derivatives to 0: The ML estimate for s is $\hat{s} = \arg\max_{s \in [0,1]} f(s)$, which occurs when $\frac{df}{ds} = 0$. Using the hint, we see that

$$f(s) = \underbrace{\left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\}\right]}_{A} \log s + \underbrace{\left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“ham ”}\}\right]}_{B} \log(1-s)$$

has derivative equal to 0 when

$$\hat{s} = \frac{A}{A+B} = \frac{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\}}{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\} + \sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“ham ”}\}} = \frac{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam ”}\}}{n}.$$

This result is intuitive — it's the number of emails labeled “spam" divided by the total number of emails.

The ML estimate for p_j is $\hat{p}_j = \arg\max_{p_j \in [0,1]} g_j(p_j)$, which occurs when $\frac{dg_j}{dp_j} = 0$. Again using the hint, we see that

$$g_j(p_j) = \underbrace{\left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\} y_j^{(i)}\right]}_{A} \log p_j + \underbrace{\left[\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\}(1 - y_j^{(i)})\right]}_{B} \log(1 - p_j)$$

has derivative equal to 0 when

$$\begin{aligned}\hat{p}_j &= \frac{A}{A+B} \\ &= \frac{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\} y_j^{(i)}}{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\} y_j^{(i)} + \sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\}(1 - y_j^{(i)})} \\ &= \frac{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\} y_j^{(i)}}{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{``ham''}\}}.\end{aligned}$$

This result is also intuitive — it's the number of times word j occurred in an email labeled “ham" divided by the total number of emails labeled “ham".

Finally, by pattern-matching, the ML estimate for q_j is

$$\hat{q}_j = \frac{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam”}\} y_j^{(i)}}{\sum_{i=1}^{n} \mathbf{1}\{c^{(i)} = \text{“spam”}\}}.$$

Wonderful, now we can write up an algorithm that computes all those ML estimates above. Once we learn the parameters θ, we can treat them as fixed and start doing prediction.

Discussion

Topic: Parameter Learning - Naive Bayes Classification / The Naive Bayes Classifier: Training

Show Discussion