Bookmarks

- Introduction
- Part 1: Probability and Inference
- Part 2: Inference in Graphical Models
- **Part 3: Learning Probabilistic Models**

Week 8: Introduction to Learning Probabilistic Models

Week 8: Introduction to Parameter Learning - Maximum Likelihood and MAP Estimation
Exercises due Nov 10, 2016 at 01:30 IST

Week 8: Homework 6
Homework due Nov 10, 2016 at 01:30 IST

Part 3: Learning Probabilistic Models > Week 9: Parameter Learning - Naive Bayes Classification > The Naive Bayes Classifier: Prediction

The Naive Bayes Classifier: Prediction

Bookmark this page

THE NAIVE BAYES CLASSIFIER: PREDICTION (PREFACE)

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

Practice problem: Once we learn the parameters θ, we can treat them as fixed and start doing prediction. Let's now look at classifying whether a new email that's not in our training data is spam or ham. This new email has random, unobserved label C, which we would like to infer, but we only get to see its features $Y_1 = y_1, Y_2 = y_2, \ldots, Y_J = y_J$. Assuming that θ is known and fixed, figure out what the MAP estimate for label C is given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_J = y_J$.

The Naive Bayes Classifier: Prediction

0:29 / 0:00 1.0x

email

and with this theta we've already learned during the training phase.

So theta is treated as fixed constant.

And so now we're interested in, well,

we see a new email, which we can extract these features,

Y 1 up to Y J for.

And in particular, we observe these little y 1 up to y J.

And we want to know whether this new email is spam or not.

So we're interested in, well, what is the probability that C

is spam given these features?

Well what is this equal to?

We'll just going to use Bayes' rule.

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: PREDICTION (COURSE NOTES)

Unleashing Bayes' rule,

$$
\begin{aligned}
&p_{C|Y_1,\ldots,Y_J}(\text{``spam''} \,|y_1,\ldots,y_J) \\
&= \frac{p_C(\text{``spam''})p_{Y_1,\ldots,Y_J|C}(y_1,\ldots,y_J|\text{``spam''})}{p_{Y_1,\ldots,Y_J}(y_1,\ldots,y_J)} \\
&= \frac{p_C(\text{``spam''})p_{Y_1,\ldots,Y_J|C}(y_1,\ldots,y_J|\text{``spam''})}{p_C(\text{``spam''})p_{Y_1,\ldots,Y_J|C}(y_1,\ldots,y_J|\text{``spam''}) + p_C(\text{``ham''})p_{Y_1,\ldots,Y_J|C}(y_1,\ldots,y_J|\text{``ham''})} \\
&= \frac{p_C(\text{``spam''})\prod_{j=1}^J p_{Y_j|X}(y_j|\text{``spam''})}{p_C(\text{``spam''})\prod_{j=1}^J p_{Y_j|C}(y_j|\text{``spam''}) + p_C(\text{``ham''})\prod_{j=1}^J p_{Y_j|X}(y_j|\text{``ham''})} \\
&= \frac{s\prod_{j=1}^J q_j^{y_j}(1-q_j)^{1-y_j}}{s\prod_{j=1}^J q_j^{y_j}(1-q_j)^{1-y_j} + (1-s)\prod_{j=1}^J p_j^{y_j}(1-p_j)^{1-y_j}},
\end{aligned}
$$

where for simplicity we've dropped the hats on the parameters even though the parameter values we use are estimated from training data.

Of course,

$$
\begin{aligned}
&p_{C|Y_1,\ldots,Y_J}(\text{``ham''} \,|y_1,\ldots,y_J) \\
&= 1 - p_{C|Y_1,\ldots,Y_J}(\text{``spam''} \,|y_1,\ldots,y_J) \\
&= \frac{(1-s)\prod_{j=1}^J p_j^{y_j}(1-p_j)^{1-y_j}}{s\prod_{j=1}^J q_j^{y_j}(1-q_j)^{1-y_j} + (1-s)\prod_{j=1}^J p_j^{y_j}(1-p_j)^{1-y_j}}.
\end{aligned}
$$

The MAP estimate for C is

$$
\widehat{C}_{\text{MAP}} = \begin{cases} \text{``spam''} & \text{if } p_{C|Y_1,\ldots,Y_J}(\text{``spam''} \,|y_1,\ldots,y_J) \geq p_{C|Y_1,\ldots,Y_J}(\text{``ham''} \,|y_1,\ldots,y_J) \\ \text{``ham''} & \text{otherwise.} \end{cases}
$$

Note that here we're breaking ties in favor of spam. The above is equivalent to looking at whether the *odds ratio*

$$\frac{p_{C|Y_1,...,Y_J}(\text{“spam ”}|y_1,\dots,y_J)}{p_{C|Y_1,...,Y_J}(\text{“ham ”}|y_1,\dots,y_J)}$$

is at least 1, or whether the *log odds ratio*

$$\log\frac{p_{C|Y_1,...,Y_J}(\text{“spam ”}|y_1,\dots,y_J)}{p_{C|Y_1,...,Y_J}(\text{“ham ”}|y_1,\dots,y_J)}$$

is at least 0. In practice the log odds ratio can be much more numerically stable to compute since, pushing in the log, we end up taking sums and differences of log probabilities rather than multiplying a large number of probabilities.

Discussion

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

Show Discussion