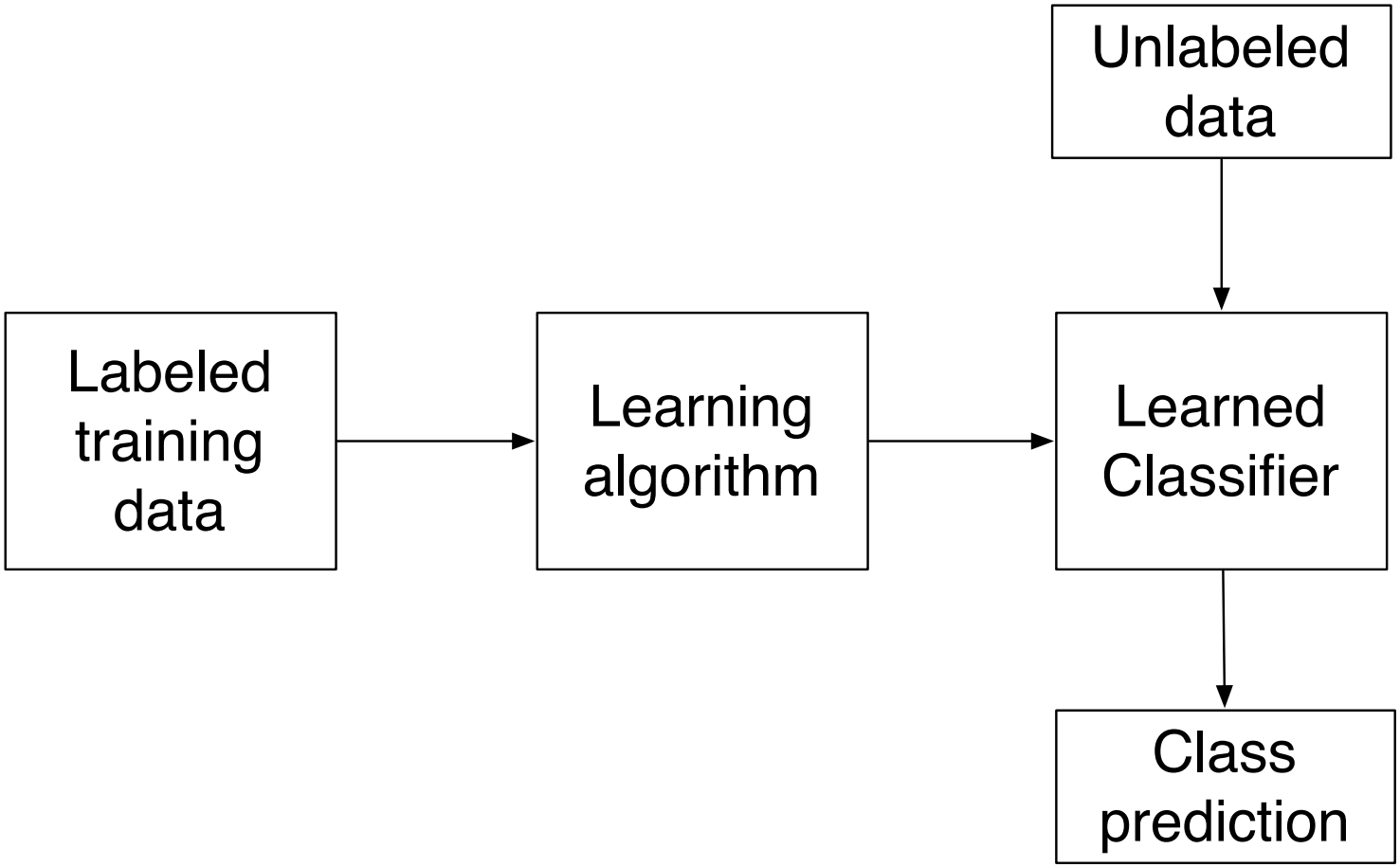
Naive Bayes

David M. Blei

COS424
Princeton University

February 14, 2008

Classification



The next few lectures are about *classification*.

Binary text classification

- The data are D documents and their classes $\{w_{d,1:N}, c_d\}$, e.g., emails and whether they are spam.
- Goal: build a classifier that can predict the category of the email.
- Divide the data into a *training set* and *testing set*.
- We are allowed to use the labels of the training set to build the classifier; we can only use the test labels to evaluate it.
- One good *evaluation metric* is *accuracy*

$$\frac{\#\text{ correctly predicted labels in the test set}}{\#\text{ of instances in the test set}}$$

Naive Bayes model

- Define a joint distribution over words and classes

$$p(c, w_{1:N} \mid \pi, \theta) = p(c \mid \pi) \prod_{n=1}^{N} p(w_n \mid \theta_c)$$

- The parameters of the model are
  - Class probabilities π :
    a distribution over classes, e.g., "spam" or "ham."
  - Class conditional probabilities θ_c :
    the conditional probability table of words given class.
- Q: What are the sizes of these parameters?
- Q: Given a model, how do we classify?
- Q: What are the independence assumptions behind this model?

Class prediction with naive Bayes

- π is a Bernoulli parameter (for binary classification);
  θ_c is a V-dimensional distribution
- We classify using the *posterior* distribution over classes given the words of the (unlabeled) document:

$$p(c \mid w_{1:N}, \pi, \theta) \propto p(c \mid \pi) \prod_{n=1}^{N} p(w_n \mid \theta_c)$$

- This model assumes that the words are conditionally independent given the class.
- Note: we do not need to compute the full posterior to classify a new data point because we are only comparing the two probabilities.

Fitting a naive Bayes model with maximum likelihood

- We compute the maximum likelihood estimate of the model.
- Given data, $\{w_{d,1:N}, c_d\}_{d=1}^{D}$, the likelihood under the model is:

$$
\begin{aligned}
p(\mathcal{D} \mid \theta_{1:C}, \pi) &= \prod_{d=1}^{D} \left(p(c_d \mid \pi) \prod_{n=1}^{N} p(w_n \mid \theta_{c_d}) \right) \\
&= \prod_{d=1}^{D} \prod_{c=1}^{C} \left(\pi_c \prod_{n=1}^{N} \prod_{v=1}^{V} \theta_{c,v}^{1(w_{d,n}=v)} \right)^{1(c_d=c)}
\end{aligned}
$$

- Take logs:

$$
\begin{aligned}
\mathcal{L}(\pi, \theta_{1:C}; \mathcal{D}) = &\sum_{d=1}^{D} \sum_{c=1}^{C} 1(c_d = c) \log \pi_c + \\
&\sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{v=1}^{V} 1(c_d = c) 1(w_{d,n} = v) \log \theta_{c,v}
\end{aligned}
$$

MLE (cont)

- The log likelihood decomposes into two simpler likelihoods.
- For the class probabilities

$$\hat{\pi}_c = \frac{n_c}{D}$$

- For the class conditional distributions

$$\hat{\theta}_{c,w} = \frac{n_{c,w}}{\sum_{w'} n_{c,w'}}$$

- This procedure is intuitive!

Full procedure

- Estimate the model from the training set.
- Predict the class of each test example.
- Compute the accuracy.

Naive Bayes case study

- Training set: 10,000 emails that are either SPAM or HAM
- Testing set: 1,000 additional emails
- Train a Naive Bayes classifier on (a subset of) the training set
- Predict SPAM/HAM on the test set and compute accuracy.

Mark – I am working with the East power desk to purchase space for an EnronOnline banner ad on a PJM website. We are buying 7 ads at 500/month/ad for 3 months ($10,500 total). They are running this ad as a pilot program offered for only 3 months. I am attaching the agreement they sent to us. I would like to revise section 2.01 to state that EnronOnline has first right of refusal to keep the ad on their site if they extend the program after three months. Could you help me revise this agreement?
Thanks
Kal

Mark – I am working with the East power desk to purchase space for an EnronOnline banner ad on a PJM website. We are buying 7 ads at 500/month/ad for 3 months ($10,500 total). They are running this ad as a pilot program offered for only 3 months. I am attaching the agreement they sent to us. I would like to revise section 2.01 to state that EnronOnline has first right of refusal to keep the ad on their site if they extend the program after three months. Could you help me revise this agreement?
Thanks
Kal

HAM!

Body Wrap at Home to lose 6-20 inches in one hour. With Bodywrap we guarantee: you'll lose 6-8 Inches in one hour 100% Satisfaction or your money back¡BR¿¡/P¿ Bodywrap is soothing formula that contours, cleanses and rejuvenates your body while reducing inches.¡BR¿ ambuscade eunice diffeomorphism sycamore kampala excelled possessor dobbin aqueduct tertiary smudgy beebread shawnee flat anybody multi necromancy harriet seder amherst paleozoic jejune irredentism cornet buckley eleanor casteth ponce administrate babysitter admittance abernathy bethesda busy joaquin casebook unidimensional carboloy captious bracelet anniversary edwin albumin tangent

Body Wrap at Home to lose 6-20 inches in one hour. With Bodywrap we guarantee: you'll lose 6-8 Inches in one hour 100% Satisfaction or your money back¡BR¿¡/P¿ Bodywrap is soothing formula that contours, cleanses and rejuvenates your body while reducing inches.¡BR¿ ambuscade eunice diffeomorphism sycamore kampala excelled possessor dobbin aqueduct tertiary smudgy beebread shawnee flat anybody multi necromancy harriet seder amherst paleozoic jejune irredentism cornet buckley eleanor casteth ponce administrate babysitter admittance abernathy bethesda busy joaquin casebook unidimensional carboloy captious bracelet anniversary edwin albumin tangent

SPAM!

HAM words

enron	8.58508e+00
scott	6.50723e+00
chris	6.43892e+00
edison	6.13924e+00
jeff	6.10057e+00
disclosure	5.97333e+00
mw	5.94861e+00
pge	5.92610e+00
karen	5.89284e+00
kimberly	5.82908e+00

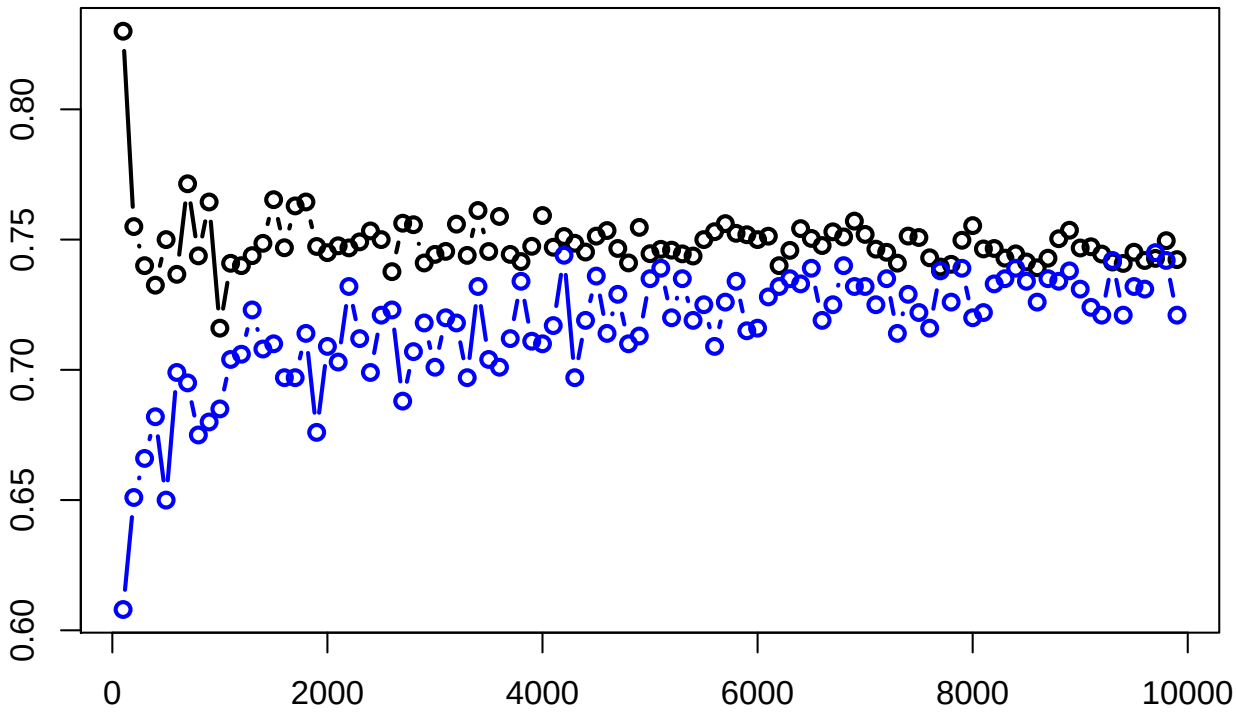
SPAM words

taacaeeccorpenroncom	8.14474e+00
ur	7.80475e+00
contentdtexthtml	7.50449e+00
multipart	7.11542e+00
nds	7.10469e+00
ger	7.10006e+00
thr	7.10006e+00
reas	7.09384e+00
bgcolordffffff	7.05898e+00
tdtd	7.01361e+00

More SPAM words

bilion	6.51536e+00
namedgenerator	6.44339e+00
tras	6.40845e+00
illustrator	6.36260e+00
contentdmshtml	6.20141e+00
meds	6.18801e+00
wastes	6.15868e+00
omit	6.14268e+00
pills	6.02968e+00
spe	5.99834e+00
mime	5.99445e+00

Sensitivity to training size



(black = training accuracy; blue = test accuracy)

A problem...

Suppose we see a rare word like "peanut" in one of our SPAM emails?

- Q: What will $\theta_{\text{spam,peanut}}$ be?
- Q: What will $\theta_{\text{ham,peanut}}$ be?

A problem...

Suppose we see a rare word like "peanut" in one of our SPAM emails?

- Q: What will $\theta_{\text{spam,peanut}}$ be?
- Q: What will $\theta_{\text{ham,peanut}}$ be?

This is what happens:

- A: $\theta_{\text{spam,peanut}}$ will be something.
- A: $\theta_{\text{ham,peanut}}$ will be 0.

Is this reasonable?

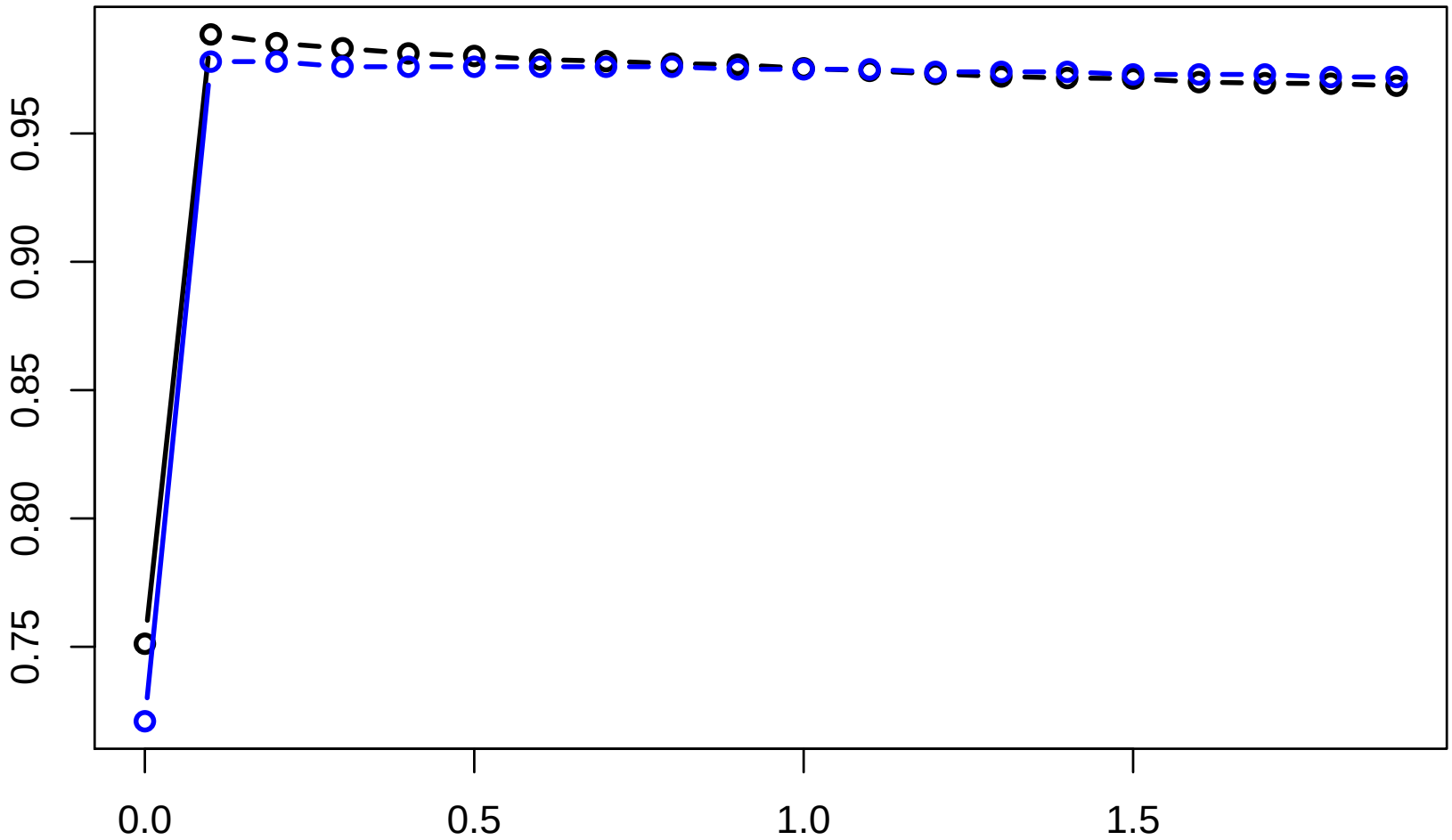
Smoothing

- In *smoothing* we add a number λ to the observed per-word class counts, including the 0-counts.
- Thus,

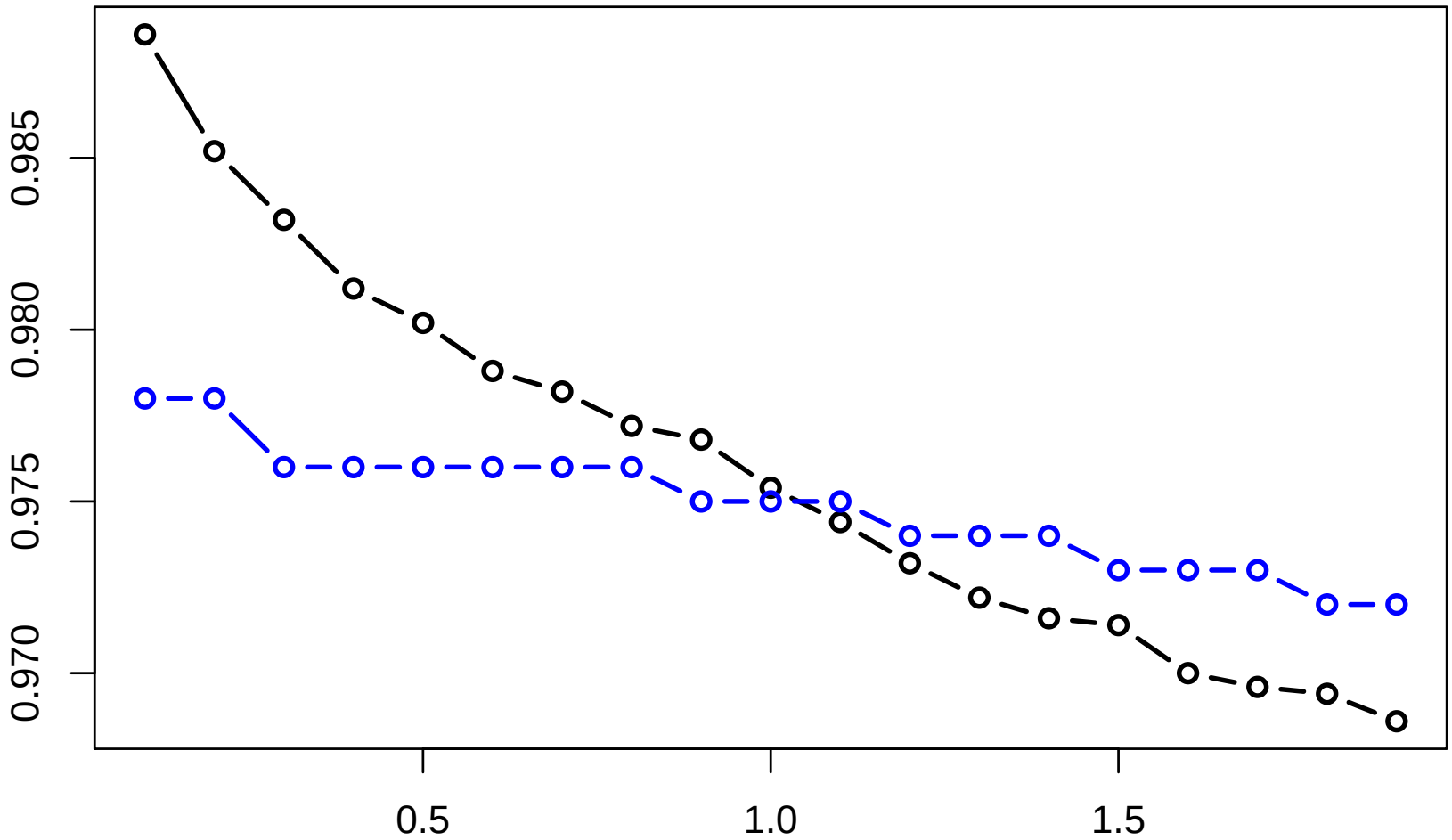
$$\hat{\theta}_{c,w} = \frac{n_{c,w} + \lambda}{\sum_{w'} n_{c,w'} + V\lambda}$$

- This upweights the rare words to something non-zero and downweights (a little, usually) the frequent words.
- Terminology
  - $\lambda = 1$: Laplace smoothing
  - $\lambda = 0.5$: Jeffrey's smoothing

Sensitivity to smoothing



Sensitivity to smoothing



SPAM words (0.1 smoothing)

gbbl	1.05488e+01
widthd	9.83269e+00
heightd	9.64469e+00
borderd	9.40989e+00
geec	9.02820e+00
cellpaddingd	8.96986e+00
voip	8.87144e+00
cellspacingd	8.86078e+00
hotfix	8.77111e+00
ur	8.60916e+00

HAM words (0.1 smoothing)

ferc	7.82131e+00
enrons	7.60930e+00
scott	7.45650e+00
pipeline	7.33990e+00
chris	7.29062e+00
enron	7.18227e+00
ena	7.13472e+00
joe	7.07833e+00
yards	6.96004e+00

Questions about NB

- What is strange about the NB model of text? Is it correct?
- What effect do the assumptions have on this classifier?
- Can you adapt NB to different data, e.g., vectors of reals? How?