# Week 3

# Agenda

1. Calendar
    - Combining week 9 and week 10.  Doesn’t change due dates.
    - Backprop deep dive during office hour on Oct 30th.  Will be recorded.
2. Async Review
3. Graham article discussion
4. Norvig article discussion
5. Domingos paper discussion
6. Time-permitting: Notebooks

For next week: Start project 1

# Quizzes as Warm-up

What makes naive Bayes "naive"?

It usually doesn't work very well.

The assumption that the classes are independent.

The assumption that the features are independent given the class.

Feature selection is important because

It can remove poorly estimated features.

It keeps only the features that give the optimal performance.

Summing log probabilities is equivalent to multiplying probabilities.

True

False

A perfectly calibrated classifier is ____% accurate on examples where the posterior probability is 85%.

P(spam) = 0.4 P("viagra") = 0.05 P("viagra" | spam) = 0.06

P(spam|"viagra") = ?

What is the Laplace (with k=1) smoothed estimate for P(sun) given this data: domain: {sun,rain,wind} observations: [sun,rain,rain,wind,sun,sun]

In our one-feature spam classifier, we have made no assumptions of independence.

True

False

If A and B are conditionally independent given C, then A and B are independent.

True

False

If A and B are independent, then A and B are conditionally independent given C?

True

False

# Naive Bayes

## Bayes's Rule

- Update our belief about X, given evidence E.

$$\begin{aligned} P(X|E) &= P(X, E) / P(E) \text{ (apply the definition)} \\ &= P(E, X) / P(E) \text{ (reorder the variables)} \\ &= P(X)\, P(E|X) / P(E) \text{ (apply the definition)} \end{aligned}$$

- Terminology:
  - **Prior**: $P(X)$
  - **Posterior**: $P(X|E)$
  - **Likelihood**: $P(E|X)$
- Why is Bayes's rule helpful?
  - Often one conditional is easier to come by than the other.

1. How long does it take to train? Can training be parallelized?
2. How fast is the trained model at making predictions?
3. Why can Naive Bayes be thought of as an 'online model'?

## Parameter Estimation

- Where do the probabilities come from?
  - Ask an expert? People are bad at estimating probabilities.
- Use training data and maximum likelihood estimation (MLE).
- **Likelihood**: probability of observed features, given parameters.
- Let $\Theta$ represent our set of parameters.
- We want the $\Theta$ that maximizes the training data likelihood:

$$\begin{aligned} \Theta_{ML} &= \text{argmax}_\Theta \prod_j P_\Theta(X_j | Y_j) \\ &= \text{argmax}_\Theta \prod_j \prod_i P_\Theta(F_{ji} | Y_j) \\ &= \text{argmax}_\Theta \sum_j \sum_i \log P_\Theta(F_{ji} | Y_j) \end{aligned}$$

# Feature Engineering

Video Slide Presentation

## How to Select Features

- Choose a vocabulary by:
  - Frequency
  - Odds ratio: $P(x|spam) / P(x|ham)$
  - Information gain
- Deal with text feature variations:
  - Tokenization (he'll → he 'll)
  - Casing (use standard form)
  - Stemming (jumped → jump; went → go)
  - Other normalizations (numbers, dates, etc.)

1. Why does feature engineering matter? What makes a feature a 'good' feature?
2. What is feature selection? Why does it matter with Naive Bayes?

# Spam Classification

## Naive Bayes for Spam

- Here's our model:
  - Wi is the word at position i in the input.
  - $P(Y|X) \sim P(Y)\,P(W_1|Y)\,P(W_2|Y) \ldots P(W_n|Y)$
- "Bag of Words" (BOW) assumption:
  - Usually, each feature has its own distribution: $P(F_i|Y)$
  - Here, each position has the same distribution: $P(W|Y)$
  - Keeps the number of parameters manageable.

## Spam Classification Example

| Feature | P(f\|spam) | P(f\|ham) | Total Spam | Total Ham |
|---|---|---|---|---|
| (prior) | 0.4000 | 0.6000 | –0.92 | –0.51 |
| dear | 0.0013 | 0.0009 | –7.56 | –7.52 |
| sir | 0.0023 | 0.0004 | –13.64 | –15.35 |
| , | 0.0220 | 0.0241 | –17.45 | –19.07 |
| first | 0.0018 | 0.0023 | –23.77 | –25.15 |
| I | 0.0062 | 0.0119 | –28.86 | –29.57 |
| must | 0.0034 | 0.0028 | –34.54 | –35.45 |
| solicit | 0.0007 | 0.0002 | –41.08 | –43.97 |

- We use log probabilities to prevent underflow.
- $P(\text{spam}|X) = 0.95$
- $P(\text{ham}|X) = 0.05$
- $\text{prediction} = \text{argmax}_y \log P(y) + \sum_i \log P(F_i|Y)$

1. Why do you think Naive Bayes might be so closely associated with text classification? Or put another way, what about NB makes it well suited to working with text based features?
2. Where do the training labels come from in a commercial spam filter?

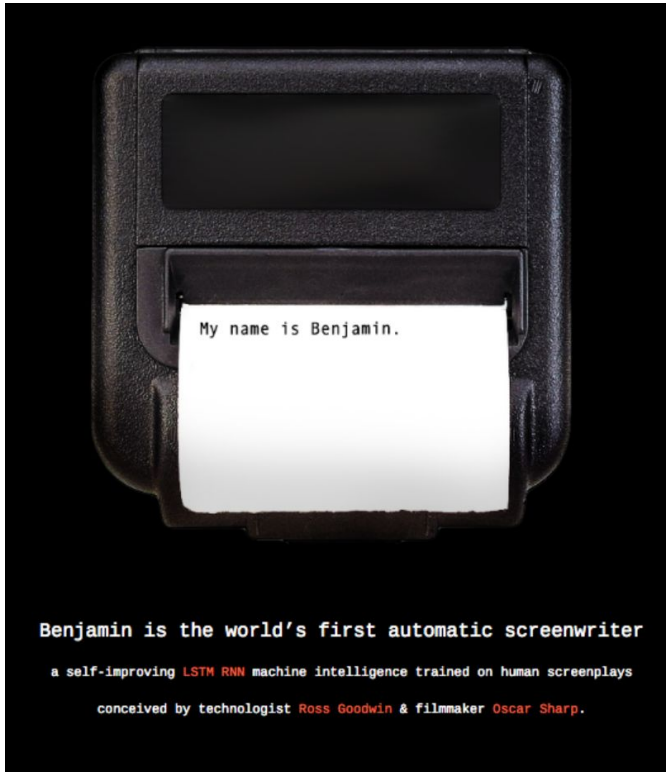
# Generative Modeling

## Generative Story for Naive Bayes

- Naive Bayes is a generative model:
  - $P(Y|X) \sim P(Y)P(W_1|Y)P(W_2|Y)...P(W_n|Y)$
- To generate a document:
  - Pick a class spam/ham according to $P(Y)$.
  - Repeat until you have enough words:
    - Pick a word according to $P(W|Y)$.
- Note: Not all models are generative.
  - E.g., logistic regression: not a generative model; it models posterior distribution $P(Y|X)$ directly.

1. What does it mean to be a generative model? Why isn't KNN a generative model?
2. If you generate emails with a NB spam detector, what might those emails look like?

# Generative Modeling




What do you think are the limits of generative modeling today?

https://distill.pub/2016/handwriting/

https://www.youtube.com/watch?v=LY7x2Ihqjmc
https://www.youtube.com/watch?v=5qPgG98_CQ8

# Graham Article

(http://www.paulgraham.com/spam.html)

1. What are the classes?  What are the features?  What is the evaluation?
2. What are some of the engineering ‘hacks’ Graham makes?
3. How fast would his model be to train?  Predict?  Retrain with addition of one email to training data?
4. What do stemming and tokenization accomplish from a machine learning perspective?
5. Why do you think Naive Bayes is a popular (baseline) choice for doing text classification?
6. Is spam filtering (necessarily) text classification? (i.e. you build a spam filter without using text as features)?
7. What might be some non-textual features that provide evidence of spam?

# Norvig

(http://norvig.com/spell-correct.html)

1. What are the classes?  What are the features?  What is the evaluation measure?
2. How fast would his model be to train?  Predict?
3. What alternatives are there to using edit distance as evidence of P(w|c)?
4. Articles such as this one (and the Graham article) have played a role in popularizing machine learning among engineers.  Discuss the pro's and con's of diffusion in this manner?
5. If you were going to train a NB classifier to filter fraudulent rental listings from Craigslist, how might you construct it? How well would you expect it to work?

# Final Thoughts?