Naïve Bayes Classifiers

Doug Downey

Northwestern EECS 349 Spring 2015

Naïve Bayes Classifiers

- Combines all ideas we've covered
  - Conditional Independence
  - Bayes' Rule
  - Statistical Estimation
- …in a simple, yet accurate classifier
  - Classifier: Function f(***x***) from ***X*** = {<x_1, …, x_d>} to *Class*
  - E.g., ***X*** = {<GRE, GPA, Letters>}, *Class* = {yes, no, wait}

Probability => Classification (1 of 2)

- Classification task
  - Learn function f($\mathbf{x}$) from $\mathbf{X}$ = {<x_1, …, x_d>} to *Class*
  - Given: Examples D={($\mathbf{x}$, y)}

- Probabilistic Approach
  - Learn P(*Class* = y | $\mathbf{X} = \mathbf{x}$) from D
  - Given $\mathbf{x}$, pick the maximally probable y

Probability => Classification (2 of 2)

map estimation is simpler than bayesian

- More formally
  - $f(\mathbf{x}) = \arg\max_y P(\mathit{Class} = y \mid \mathbf{X} = \mathbf{x}, \theta_{\text{MAP}})$
  - θ_{MAP} : MAP parameters, learned from data
    - That is, parameters of $P(\mathit{Class} = y \mid \mathbf{X} = \mathbf{x})$
  - …we'll focus on using MAP estimate, but can also use ML or Bayesian

- Predict next coin flip? Instance of this problem
  - X = null
  - Given D= hhht…tht, estimate $P(\theta \mid D)$, find MAP
  - Predict *Class* = heads *iff* $\theta_{\text{MAP}} > ½$

first spam filters : naive bayes classifier

Example: Text Classification

Dear Sir/Madam,

We are pleased to inform you of the result of the Lottery Winners International programs held on the 30/8/2004. Your e-mail address attached to ticket number: EL-23133 with serial Number: EL-123542, batch number: 8/163/EL-35, lottery Ref number: EL-9318 and drew lucky numbers 7-1-8-36-4-22 which consequently won in the 1st category, you have therefore been approved for a lump sum pay out of US$1,500,000.00 (One Million, Five Hundred Thousand United States dollars)

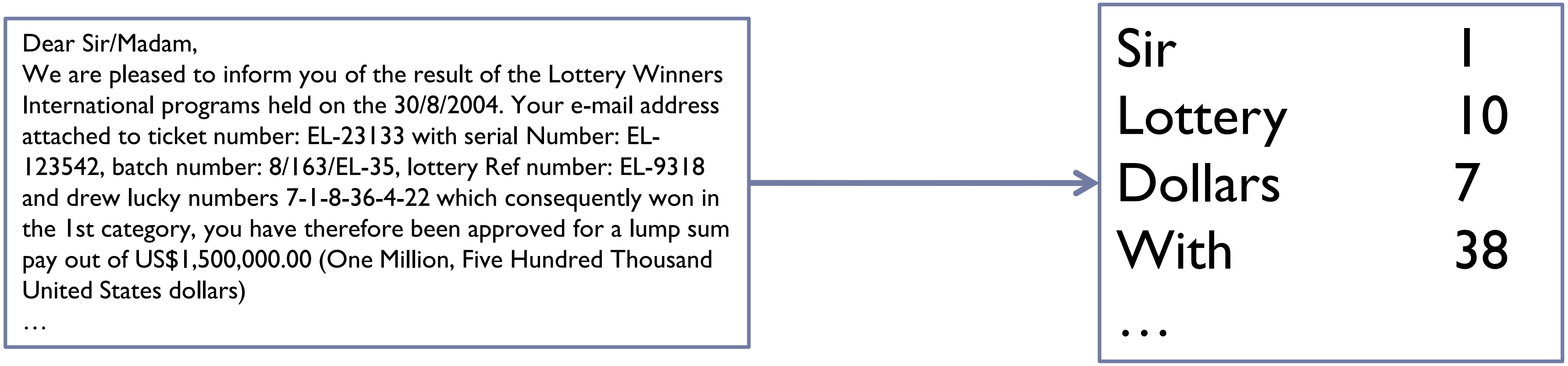
- SPAM

NOT SPAM?

not a boolean or anything,
so how do we turn it into a vector of attributes?

Representation

- ***X*** = document
- Task: Estimate P(*Class* = {spam, non-spam} | ***X***)
- Question: how to represent ***X***?
  - Lots of possibilities, common choice: "bag of words"



Dear Sir/Madam,
We are pleased to inform you of the result of the Lottery Winners International programs held on the 30/8/2004. Your e-mail address attached to ticket number: EL-23133 with serial Number: EL-123542, batch number: 8/163/EL-35, lottery Ref number: EL-9318 and drew lucky numbers 7-1-8-36-4-22 which consequently won in the 1st category, you have therefore been approved for a lump sum pay out of US$1,500,000.00 (One Million, Five Hundred Thousand United States dollars)
…

Sir	1
Lottery	10
Dollars	7
With	38
…	

Bag of Words

- Ignores Word Order, i.e.
  - No emphasis on title
  - No compositional meaning ("Cold War" -> "cold" and "war")
  - Etc.
  - But, massively reduces dimensionality/complexity
- Still and all…
  - Presence or absence of a 100,000-word vocab => 2^100,000 distinct vectors

Naïve Bayes Classifiers

- $P(Class \mid \boldsymbol{X})$ for $|Val(\boldsymbol{X})|$ = 2^100,000 requires 2^100,000 parameters
  - Problematic.
- Bayes' Rule:

  $P(Class \mid \boldsymbol{X}) = P(\boldsymbol{X} \mid Class)\, P(Class) / P(\boldsymbol{X})$
- Assume presence of word *i* is independent of all other words given *Class*:

  $P(Class \mid \boldsymbol{X}) = \Pi_i\, P(X_i \mid Class)\, P(Class) / P(\boldsymbol{X})$
- Now only 200,001 parameters for $P(Class \mid \boldsymbol{X})$

Naïve Bayes Assumption

- Features are conditionally independent given class
  - *Not* $P(\text{“Republican”}, \text{“Democrat”}) = P(\text{“Republican”})P(\text{“Democrat”})$ but instead
    $P(\text{“Republican”}, \text{“Democrat”} \mid \mathit{Class} = \text{Politics}) =$
    $P(\text{“Republican”} \mid \mathit{Class} = \text{Politics})P(\text{“Democrat”} \mid \mathit{Class} = \text{Politics})$
- Still, an absurd assumption
  - $(\text{“Lottery”} \perp \text{“Winner”} \mid \text{SPAM})$? $(\text{“lunch”} \perp \text{“noon”} \mid \text{Not SPAM})$?
- But: offers massive tractability advantages and works quite well in practice
  - Lesson: Unrealistically strong independence assumptions sometimes allow you to build an accurate model where you otherwise couldn’t

Getting the parameters from data

- Parameters $\boldsymbol{\theta} = < \theta_{ij} = P(w_i \mid Class = j) >$
- Maximum Likelihood: Estimate $P(w_i \mid Class = j)$ from *D* by counting
  - Fraction of documents in class *j* containing word *i*
  - But if word *i* never occurs in class *j* ?
- Commonly used MAP estimate:
  - $$\frac{(\#\text{ docs in class } j \text{ with word } i) + 1}{(\#\text{ docs in class } j) + 2}$$

if we have some words that never occur in training,
but if any one of those words occurs in test, prob of spam is 0
b/c of maximum likelihood estimator

Caveats

- Naïve Bayes effective as a *classifier*
- **Not** as effective in producing probability estimates
  - Π_i P(w_i | *Class*) pushes estimates toward 0 or 1
- In practice, numerical underflow is typical at classification time
  - Compare sum of logs instead of product

in the early days of spam, a thing happened:
spam detectors came out (may on naive bayes), spammers found out how it worked,
first step spammers changed their messages to use less common words like v1a instead of viagra,
but that made it easier to detect

Reading

- Elements of Statistical Learning, Ch 7:
  - http://statweb.stanford.edu/~tibs/ElemStatLearn/

other slides (unbalanced data)
acc(ZeroR) = 95%