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 $\boldsymbol{X} = \{<x_1, \ldots, x_d>\}$ to *Class*
  - E.g., $\boldsymbol{X}$ = {<GRE, GPA, Letters>}, *Class* = {yes, no, wait}

Probability => Classification (1 of 2)

- Classification task
  - Learn function f($\boldsymbol{x}$) from $\boldsymbol{X} = \{<x_1, \ldots, x_d>\}$ to *Class*
  - Given: Examples $D=\{(\boldsymbol{x}, y)\}$
- Probabilistic Approach
  - Learn P(*Class* = y | $\boldsymbol{X} = \boldsymbol{x}$) from D
  - Given $\boldsymbol{x}$, pick the maximally probable y

Probability => Classification (2 of 2)

- More formally
  - $f(\boldsymbol{x}) = \arg\max_y P(Class = y \mid \boldsymbol{X} = \boldsymbol{x}, \theta_{MAP})$
  - θ_{MAP} : MAP parameters, learned from data
    - That is, parameters of $P(Class = y \mid \boldsymbol{X} = \boldsymbol{x})$
  - …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* $\theta_{MAP} > ½$

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?

Representation

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

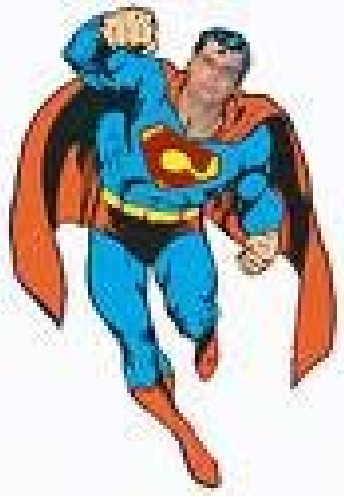
Dear Sir/Madam,
We are pleased to inform you of the result of the Lottery Winners International programs held on the 30/8/2004. Your e-mail address attached to ticket number: EL-23133 with serial Number: EL-123542, batch number: 8/163/EL-35, lottery Ref number: EL-9318 and drew lucky numbers 7-1-8-36-4-22 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*("Republican","Democrat") = *P*("Republican")*P*("Democrat")
    but instead
    P("Republican","Democrat" | *Class* = Politics) =
    P("Republican" | *Class* = Politics)*P*("Democrat" | *Class* = Politics)
- Still, an absurd assumption
  - ("Lottery" ⊥ "Winner" | SPAM)? ("lunch" ⊥ "noon" | 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 $\theta = < \theta_{ij} = P(w_i \mid Class = j) >$
- 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:
  - $$\frac{(\#\text{ docs in class } j \text{ with word } i) + 1}{(\#\text{ docs in class } j) + 2}$$

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

Reading

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