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
  - Bayes Nets
- …in a simple, yet accurate classifier
  - Classifier: Function $f(x)$ from $X = \{x_1, \ldots, x_d\}$ to Class
  - E.g., $X = \{\text{GRE, GPA, Letters}\}$, Class = \{yes, no, wait\}
Classification task
- Learn function $f(x)$ from $X = \{<x_1, \ldots, x_d>\}$ to Class
- Given: Examples $D = \{(x, y)\}$

Probabilistic Approach
- Learn $P(Class = y \mid X = x)$ from $D$
- Given $x$, pick the maximally probable $y$
Probability => Classification (2 of 2)

- More formally
  - \[ f(x) = \text{arg max}_y P(Class = y \mid X = x, \theta_{\text{MAP}}) \]
  - \( \theta_{\text{MAP}} \) : MAP parameters, learned from data
    - That is, parameters of \( P(Class = y \mid X = x) \)
  - …we’ll focus on using MAP estimate, but can also use ML or Bayesian

- Predict next coin flip? Instance of this problem
  - \( X = \text{null} \)
  - Given \( D = \text{hhht...tth} \), estimate \( P(\theta \mid D) \), find MAP
  - Predict \( Class = \text{heads} \) iff \( \theta_{\text{MAP}} > \frac{1}{2} \)
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)
Representation

- **X = document**
- **Task:** Estimate $P(Class = \{\text{spam, non-spam}\} \mid X)$
- **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)

...
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

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

- Bayes’ Rule:
  
  $$P(Class \mid X) = \frac{P(X \mid Class) \cdot P(Class)}{P(X)}$$

- Assume presence of word $i$ is independent of all other words given $Class$:
  
  $$P(Class \mid X) = \prod_i P(X_i \mid Class) \cdot P(Class) / P(X)$$

- Now only 200,001 parameters for $P(Class \mid X)$
Naïve Bayes Assumption

- Features are conditionally independent given class
  
  \[
  \text{Not } P(\text{"Republican","Democrat"}) = P(\text{"Republican"})P(\text{"Democrat"}) \\
  \text{but instead} \\
  P(\text{"Republican","Democrat" | Class = Politics}) = \\
  P(\text{"Republican" | Class = Politics})P(\text{"Democrat" | Class = Politics})
  \]

- Still, an absurd assumption
  
  ("Lottery" \perp "Winner" | SPAM)? ("lunch" \perp "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 $\theta = < \theta_{ij} = P(w_i | Class = j) >$
- Maximum Likelihood: Estimate $P(w_i | Class = j)$ 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:
  - $\frac{(\# \text{ docs in class } j \text{ with word } i) + 1}{(\# \text{ docs in class } j) + |V|}$
Caveats

- Naïve Bayes effective as a classifier

- **Not** as effective in producing probability estimates
  - $\Pi_i P(w_i | \text{Class})$ pushes estimates toward 0 or 1

- In practice, numerical underflow is typical at classification time
  - Compare sum of logs instead of product