Naïve Bayes Classifiers

Doug Downey Northwestern EECS 474

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 $f(\mathbf{x})$ from $\mathbf{X} = \{\langle x_1, \dots, x_d \rangle\}$ 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 $\mathbf{X} = \{\langle x_1, \dots, x_d \rangle\}$ to *Class*
- Given: Examples D={(x, y)}

Probabilistic Approach

- Learn P(Class = y | X = x) from D
- Given **x**, pick the maximally probable y

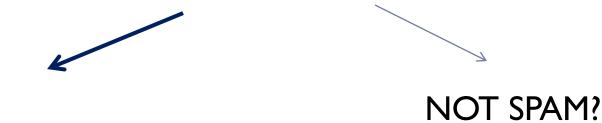
Probability => Classification (2 of 2)

- More formally
 - $f(\mathbf{x}) = \arg \max_{y} P(Class = y | \mathbf{X} = \mathbf{x}, \theta_{MAP})$
 - θ_{MAP} : MAP parameters, learned from data
 - That is, parameters of P(Class = y | 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 = null
 - Given D = hhht...tht, estimate $P(\theta \mid D)$, find MAP
 - Predict Class = heads iff $\theta_{MAP} > \frac{1}{2}$

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

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 | Sir | l I |
|---|---------|--------|
| International programs held on the 30/8/2004. Your e-mail address attached to ticket number: EL-23133 with serial Number: EL- | Lotte | ery IO |
| 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 | > Dolla | ars 7 |
| 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 | With | n 38 |
| 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

- P(Class | X) for |Val(X)| = 2^100,000 requires 2^100,000 parameters
 - Problematic.
- Bayes' Rule:
 P(Class | X) = P(X | Class) P(Class) / P(X)
- Assume presence of word *i* is independent of all other words given Class: P(Class | X) = Π_i P(X_i | Class) P(Class) / P(X)

 Now only 200,001 parameters for P(Class | 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: Overly strong independence assumptions sometimes allow you to build an accurate model where you otherwise couldn't

Getting the parameters from data

- Parameters $\theta = \langle \theta_{ij} = P(w_i | Class = j) \rangle$
- 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:
 - (# docs in class j with word i) + I (# docs in class j) + 2

Caveats

Naïve Bayes effective as a classifier

- Not as effective in producing probability estimates
 - $\Pi_i P(w_i | Class)$ pushes estimates toward 0 or 1
- In practice, numerical underflow is typical at classification time
 - Compare sum of logs instead of product