

Naïve Bayes Classifiers

Doug Downey

Northwestern EECS 395/495

Fall 2013

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

Probability => Classification (1 of 2)

- Classification Task:
 - Learn function $f(\mathbf{x})$ from $\mathbf{X} = \{<x_1, \dots, x_d>\}$ to *Class*
 - Given: Examples $D = \{(\mathbf{x}, y)\}$
- Probabilistic Approach
 - Learn $P(\text{Class} = y \mid \mathbf{X} = \mathbf{x})$ from D
 - Given \mathbf{x} , pick the maximally probable y

Probability => Classification (2 of 2)

- More formally
 - $f(\mathbf{x}) = \arg \max_y P(\text{Class} = y \mid \mathbf{X} = \mathbf{x}, \theta_{\text{MAP}})$
 - θ_{MAP} : MAP parameters, learned from data
 - That is, parameters of $P(\text{Class} = y \mid \mathbf{X} = \mathbf{x})$
 - ...we'll focus on using MAP estimate, but can also use ML or Bayesian
- Predict next coin flip? Instance of this problem
 - $X = \text{null}$
 - Given $D = \text{hhht...tth}$, estimate $P(\theta \mid D)$, find MAP
 - Predict $\text{Class} = \text{heads}$ iff $\theta_{\text{MAP}} > \frac{1}{2}$

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

NOT SPAM?

Representation

- X = document
- Estimate $P(\text{Class} = \{\text{spam}, \text{non-spam}\} \mid X)$
- Question: how to represent X ?
 - One dimension for each possible e-mail, i.e. possible permutation of words?
 - No.
 - 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...
 - Recording presence or absence of a 100,000-word vocab entails $2^{100,000}$ distinct vectors

Naïve Bayes Classifiers



- $P(\text{Class} \mid \mathbf{X})$ for $|\text{Val}(\mathbf{X})| = 2^{100,000}$ requires $2^{100,000}$ parameters
 - Problematic.
- Bayes' Rule:
$$P(\text{Class} \mid \mathbf{X}) = P(\mathbf{X} \mid \text{Class}) P(\text{Class}) / P(\mathbf{X})$$
- Assume presence of word i is independent of all other words given Class :
$$P(\text{Class} \mid \mathbf{X}) = \prod_i P(w_i \mid \text{Class}) P(\text{Class}) / P(\mathbf{X})$$
- Now only 200,001 parameters for $P(\text{Class} \mid \mathbf{X})$

Naïve Bayes Assumption

- Features are conditionally independent given class
 - *Not* $P(\text{"Republican"}, \text{"Democrat"}) = P(\text{"Republican"})P(\text{"Democrat"})$
but instead
$$P(\text{"Republican"}, \text{"Democrat"} \mid \text{Class} = \text{Politics}) =$$
$$P(\text{"Republican"} \mid \text{Class} = \text{Politics})P(\text{"Democrat"} \mid \text{Class} = \text{Politics})$$
- 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: Overly strong independence assumptions sometimes allow you to build an accurate model where you otherwise couldn’t

Getting the parameters from data

- Parameters $\theta = \langle \theta_{ij} = P(w_i | \text{Class} = j) \rangle$
- Maximum Likelihood: Estimate $P(w_i | \text{Class} = j)$ 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
 - $\prod_i P(w_i | Class)$ pushes estimates toward 0 or 1
- In practice, numerical underflow is typical at classification time
 - Compare sum of logs instead of product

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: *Burglar, Earthquake, Alarm, JohnCalls, MaryCalls*