Naïve Bayes Classifiers

Doug Downey

Northwestern EECS 395/495: Special Topics in Machine Learning

Winter 2010

Naïve Bayes Classifiers

- Combines all ideas we've covered
  - Conditional Independence
  - Bayes' Rule
  - Statistical Estimation
  - Machine Learning
- …in a simple, yet empirically powerful classifier
  - Classifier: Function f(***x***) from ***X*** = {<x1, …, xd>} to *Class*
  - E.g., ***X*** = {<GRE, GPA, Letters>}, *Class* = {yes, no, wait}

Probability => Classification (1 of 2)

- Classification Task:
  - Learn function f(***x***) from ***X*** = {<x1, …, xd>} 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
  - $f(\boldsymbol{x}) = \arg\max_y P(Class = y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\theta}_{\text{MAP}})$
  - $\boldsymbol{\theta}_{\text{MAP}}$: 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 $\theta_{\text{MAP}} > ½$

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
- Estimate P(*Class* = {spam, non-spam} | ***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

- $P(Class \mid \boldsymbol{X})$ for $|\mathrm{Val}(\boldsymbol{X})| = 2^{100{,}000}$ requires $2^{100{,}000}$ parameters
  – Problematic.
- Bayes' Rule:
  $P(Class \mid \boldsymbol{X}) = \mathrm{P}(\boldsymbol{X} \mid Class)\, \mathrm{P}(Class) / \mathrm{P}(\boldsymbol{X})$
- Assume presence of word *i* is independent of all other words given *Class*:
  $P(Class \mid \boldsymbol{X}) = \prod_i \mathrm{P}(w_i \mid Class)\, \mathrm{P}(Class) / \mathrm{P}(\boldsymbol{X})$
- 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)
- Generally an absurd assumption
  - (“Lottery”, “Winner” $\perp$ SPAM)? (“lunch”, “noon” $\perp$ Not SPAM)?
- But: offers massive tractability advantages and works quite well in practice
  - Lesson: Overly strong independence assumptions can be okay, and sometimes allow you to build a model where you otherwise couldn’t

Getting the parameters from data

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

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