Bayesian Methods: Naïve Bayes

Nicholas Ruozzi
University of Texas at Dallas

based on the slides of Vibhav Gogate

Last Time

- Parameter learning
  - Learning the parameter of a simple coin flipping model
- Prior distributions
- Posterior distributions
- Today: more parameter learning and naïve Bayes

Maximum Likelihood Estimation (MLE)

- **Data:** Observed set of α_H heads and α_T tails
- **Hypothesis:** Coin flips follow a binomial distribution
- **Learning:** Find the "best" θ
- **MLE:** Choose θ to maximize the likelihood (probability of D given θ)

$$\theta_{MLE} = \arg\max_{\theta} p(D|\theta)$$

MAP Estimation

- Choosing θ to maximize the posterior distribution is called maximum a posteriori (MAP) estimation

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D)$$

- The only difference between θ_{MLE} and θ_{MAP} is that one assumes a uniform prior (MLE) and the other allows an arbitrary prior

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound (again!)
  - Suppose $Y_1, \ldots, Y_N$ are i.i.d. random variables taking values in $\{0, 1\}$ such that $E_p[Y_i] = y$. For $\epsilon > 0$,

$$p\left(\left|y - \frac{1}{N}\sum_i Y_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound (again!)

  – For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p\left(\left|\theta_{true} - \frac{1}{N}\sum_i X_i\right| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

- Can use Chernoff bound (again!)

  – For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

Sample Complexity

- How many coin flips do we need in order to guarantee that our learned parameter does not differ too much from the true parameter (with high probability)?

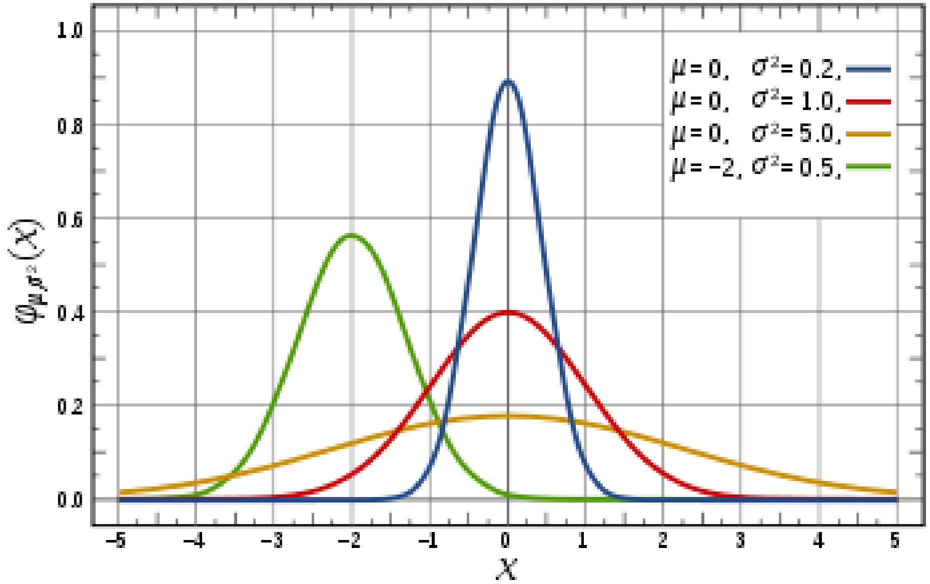
- Can use Chernoff bound (again!)
  - For the coin flipping problem with $X_1, \ldots, X_n$ iid coin flips and $\epsilon > 0$,

$$p(|\theta_{true} - \theta_{MLE}| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

$$\delta \geq 2e^{-2N\epsilon^2} \Rightarrow N \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$$

MLE for Gaussian Distributions

- Two parameter distribution characterized by a mean and a variance



$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2)$
  - $Z = X + Y \Rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- Easy to differentiate, as we will see soon!

Learning a Gaussian

- Collect data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: μ
  - Variance: σ

i	Exam Score
0	85
1	95
2	100
3	12
…	…
99	89

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian:

- **Probability of N i.i.d. samples $D = x^{(1)}, \ldots, x^{(N)}$**

$$p(D|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{i=1}^{N} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \underset{\mu,\sigma}{\arg\max}\, P(\mathcal{D} \mid \mu, \sigma)$$

- **Log-likelihood of the data**

$$\ln p(D|\mu, \sigma) = -\frac{N}{2} \ln 2\pi\sigma^2 - \sum_{i=1}^{N} \frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}$$

MLE for the Mean of a Gaussian

$$\frac{\partial}{\partial \mu} \ln p(D|\mu,\sigma) = \frac{\partial}{\partial \mu}\left[-\frac{N}{2}\ln 2\pi\sigma^2 - \sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= \frac{\partial}{\partial \mu}\left[-\sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)}{\sigma^2}$$

$$= \frac{\left[N\mu - \sum_{i=1}^{N} x^{(i)}\right]}{\sigma^2} = 0$$

$$\boxed{\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}}$$

MLE for Variance

$$\frac{\partial}{\partial \sigma} \ln p(D|\mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-\frac{N}{2}\ln 2\pi\sigma^2 - \sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{\sigma} + \frac{\partial}{\partial \sigma}\left[-\sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{\sigma} + \sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{\sigma^3} = 0$$

$$\boxed{\sigma_{MLE}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2}$$

Learning Gaussian parameters

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

$$\sigma_{MLE}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2$$

- MLE for the variance of a Gaussian is biased
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator

$$\sigma_{unbiased}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2$$

Bayesian Categorization/Classification

- Given features $x = (x_1, \dots, x_m)$ predict a label y
- If we had a joint distribution over x and y, given x we could find the label using MAP inference

$$\arg\max_y p(y|x_1, \dots, x_m)$$

- Can compute this in exactly the same way that we did before using Bayes rule:

$$p(y|x_1, \dots, x_m) = \frac{p(x_1, \dots, x_m|y)p(y)}{p(x_1, \dots, x_m)}$$

Article Classification

- Given a collection of news articles labeled by topic, goal is, given a new news article to predict its topic
  - One possible feature vector:
    - One feature for each word in the document, in order
      - x_i corresponds to the i^{th} word
      - x_i can take a different value for each word in the dictionary

Text Classification

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```

Text Classification

- $x_1, x_2, \ldots$ is sequence of words in document
- The set of all possible features (and hence $p(y|x)$) is huge
  - Article at least 1000 words, $x = (x_1, \ldots, x_{1000})$
  - x_i represents i^{th} word in document
    - Can be any word in the dictionary – at least 10,000 words
  - $10{,}000^{1000} = 10^{4000}$ possible values
  - Atoms in Universe: $\sim 10^{80}$

Bag of Words Model

- Typically assume position in document doesn't matter

$$p(X_i = x_i | Y = y) = p(X_k = x_i | Y = y)$$

  – All positions have the same distribution

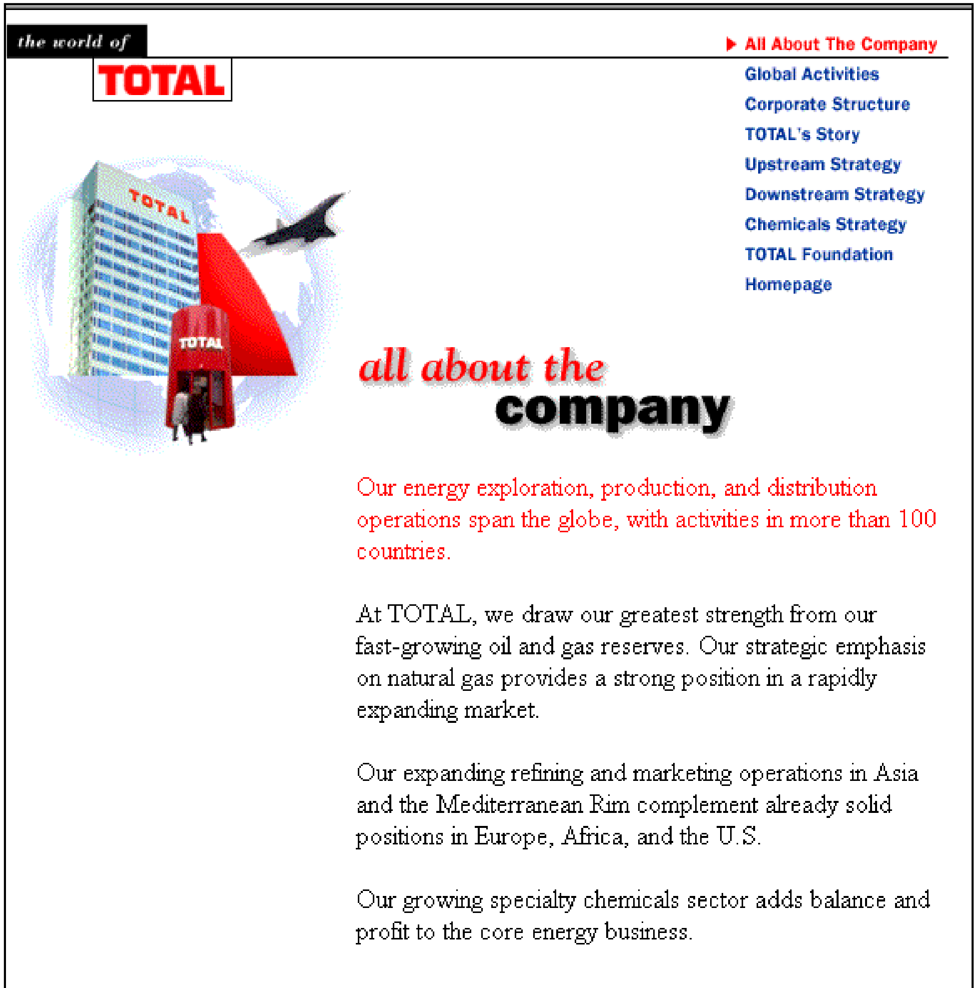
  – Ignores the order of words

  – Sounds like a bad assumption, but often works well!

- Features

  – Set of all possible words and their corresponding frequencies (number of times it occurs in the document)

Bag of Words



the world of

TOTAL

- All About The Company
- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage

all about the **company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

→

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Need to Simplify Somehow

- Even with the bag of words assumption, there are too many possible outcomes
  - Too many probabilities

$$p(x_1, \ldots, x_m | y)$$

- Can we assume some are the same?

$$p(x_1, x_2 | y) = p(x_1 | y)\, p(x_2 | y)$$

  - This is a conditional independence assumption

Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution for X is independent of the value of Y, given the value of Z

$$p(X|Y,Z) = P(X|Z)$$

- Equivalent to

$$p(X,Y|Z) = p(X|Z)P(Y|Z)$$

Naïve Bayes

- Naïve Bayes assumption
  - Features are independent given class label

$$p(x_1, x_2|y) = p(x_1|y)\, p(x_2|y)$$

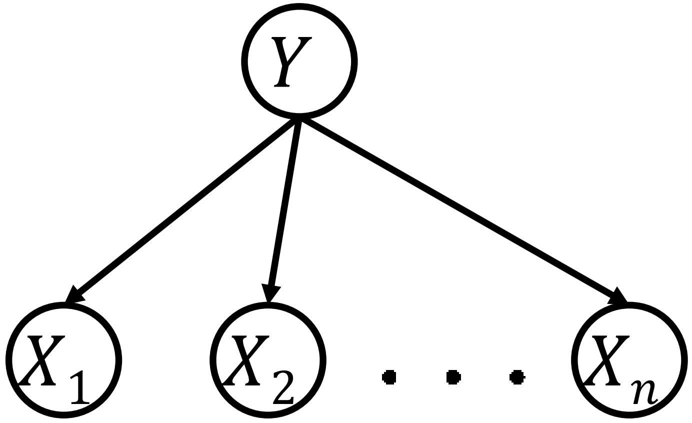
  - More generally

$$p(x_1, \ldots, x_m|y) = \prod_{i=1}^{m} p(x_i|y)$$

- How many parameters now?
  - Suppose x is composed of m binary features

The Naïve Bayes Classifier

- **Given**
  - Prior $p(y)$
  - m conditionally independent features X given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$



- **Classify via**

$$y^* = h_{NB}(x) = \arg\max_y p(y)p(x_1, \ldots, x_m|y)$$
$$= \arg\max_y p(y) \prod_i^m p(x_i|y)$$

MLE for the Parameters of NB

- **Given dataset, count occurrences for all pairs**
  - $Count(X_i = x_i, Y = y)$ is the number of samples in which $X_i = x_i$ and $Y = y$

- **MLE for discrete NB**

$$p(Y = y) = \frac{Count(Y = y)}{\sum_{y'} Count(Y = y')}$$

$$p(X_i = x_i | Y = y) = \frac{Count(X_i = x_i, Y = y)}{\sum_{x_i'} Count\,(X_i = x_i', Y = y)}$$

Naïve Bayes Calculations

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Subtleties of NB Classifier: #1

- Usually, features are not conditionally independent:

$$p(x_1, \dots, x_m | y) \neq \prod_{i=1}^{m} p(x_i | y)$$

- The naïve Bayes assumption is often violated, yet it performs surprisingly well in many cases

- Plausible reason: Only need the probability of the correct class to be the largest!
  - Example: binary classification; just need to figure out the correct side of 0.5 and not the actual probability (0.51 is the same as 0.99).

Subtleties of NB Classifier: #2

- What if you never see a training instance $(X_1 = a, Y = b)$
  - Example: you did not see the word Enlargement in spam!
  - Then $p(X_1 = a|Y = b) = 0$
  - Thus no matter what values $X_2, \cdots, X_m$ take

$$P(X_1 = \mathrm{a}, X_2 = x_2, \cdots, X_m = x_m|Y = b) = 0$$

  - Why?

Subtleties of NB Classifier: #2

- To fix this, use a prior!
  - Already saw how to do this in the coin-flipping example using the Beta distribution
  - For NB over discrete spaces, can use the Dirichlet prior
  - The Dirichlet distribution is a distribution over $z_1, \ldots, z_k \in (0,1)$ such that $z_1 + \cdots + z_k = 1$ characterized by k parameters $\alpha_1, \ldots, \alpha_k$

$$f(z_1, \ldots, z_k; \alpha_1, \ldots, \alpha_k) \propto \prod_{i=1}^{k} z_i^{\alpha_i - 1}$$

    - Called smoothing, what are the MLE estimates under these kinds of priors?