Introduction to Probability and Naïve Bayes

Sethserey Sam & Tain Ngounly

Part1: Probability

- Definition
- Probability rules
- Bayes theorem

Definitions

The *sample space* is a set S composed of all the possible outcomes of an experiment.

- If we flip a coin twice, the sample space might be taken as $S = \{hh, ht, th, tt\}$
- The sample space for the outcomes of the 2000 US presidential election might be taken as S = {*Bush, Gore, Nader, Buchanan*}

Definitions (continued)

An *event* is a subset of the sample space, or

$A \subset S$ is an "event".

- A = {*hh*} would be the event that heads occur twice when a coin is flipped twice.
- A = {*Gore, Bush*} would be the event that a major party candidate wins the election.

Probability

1) For every event $A \subset S$, $Pr(A) \geq 0$
2) $\Sigma p_i = 1$
3) All of the probabilities constitute the probability distribution.
4) For all $A \subset S$, $Pr(A) + Pr(\bar{A}) = 1$

Statistical Independence

H is <u>statistically independent</u> of G if:

$Pr(H \mid G) = Pr(H)$

Recall that $Pr(H \text{ and } G) = Pr(G)Pr(H \mid G)$

If H and G are independent, then we can replace $Pr(H \mid G)$ with $Pr(H)$

Statistical Independence

Thus for independent events,

Pr(H and G) = Pr(G)Pr(H)

Example: the probability of having a boy, girl, and then boy in a family of three

The sex of each child is independent of the sex of the others, thus we can calculate

Pr(B and G and B) = Pr(B)Pr(G)Pr(B)

Pr(B and G and B) = (1/2)(1/2)(1/2) = 1/8

Review of Probability Rules

1) $Pr(G \text{ or } H) = Pr(G) + Pr(H) - Pr(G \text{ and } H)$; for mutually exclusive events, $Pr(G \text{ and } H) = 0$
2) $Pr(G \text{ and } H) = Pr(G)Pr(H \mid G)$, also written as $Pr(H \mid G) = Pr(G \text{ and } H)/Pr(G)$
3) If G and H are independent then, $Pr(H \mid G) = Pr(H)$, thus $Pr(G \text{ and } H) = Pr(G)Pr(H)$

Bayesian Statistics

- Formal way of updating our beliefs about parameters given data that actually occurred
- Inference/hypothesis testing method
  - Is A different from B? How much do we believe this?
  - Are A and B the same?
- Parameter estimation method
  - Way of modeling how the world works
- Decision making method
  - What is the best action to take?

Applications

- Decision making
  - Medicine, management, economics

- Human-computer interactions
  - “Intelligent” software

- Modeling human decisions

- Modeling perception & cognition
  - Helmholtz
  - Human vision as bayesian inference

Bayes Theorem

Sometimes we have prior information or beliefs about the outcomes we expect.

Example: I am thinking of buying a used Saturn at Honest Ed's. I look up the record of Saturns in an auto magazine and find that, unfortunately, 30% of these cars have faulty transmissions. To get a better estimate of the particular car I want to purchase, I hire a mechanic who can make a guess on the basis of a quick drive around the block. I know that the mechanic is able to pronounce 90% of the faulty cars faulty and he is able to pronounce 80% of the good cars good.

Bayes Theorem (continued)

What is the chance that the Saturn I'm thinking of buying has a faulty transmission:

1) Before I hire the mechanic?
2) If the mechanic pronounces it faulty?
3) If the mechanic pronounces it ok?

Bayes Theorem (continued)

Using Bayes Theorem:

$$\Pr(A \mid MA) = \frac{\Pr(A)\Pr(MA \mid A)}{\Pr(A)\Pr(MA \mid A) + \Pr(\bar{A})\Pr(MA \mid \bar{A})}$$

A = transmission is actually faulty

MA= mechanic declares transmission faulty

Bayes Theorem (continued)

$Pr(A) = .3$

$Pr(MA \mid A) = .9$

$Pr(\bar{A}) = .7$

$Pr(MA \mid \bar{A}) = .2$

$$Pr(A \mid MA) = \frac{(.3)(.9)}{(.3)(.9) + (.7)(.2)} = .27/.41 = .659$$

Bayes Theorem (continued)

Interpretation: The probability that the transmission is faulty if the mechanic declares it faulty is 0.659, which is a much better estimate than our guess based on the auto magazine (0.3).

Another Example

What is the probability that the transmission is faulty if the mechanic claims it is good?

$$\Pr(A \mid M\,\bar{A}) = \frac{\Pr(A)\Pr(M\,\bar{A} \mid A)}{\Pr(A)\Pr(M\,\bar{A} \mid A) + \Pr(\bar{A})\Pr(M\,\bar{A} \mid \bar{A})}$$

$$\Pr(A \mid M\,\bar{A}) = \frac{(.3)(.1)}{(.3)(.1) + (.7)(.8)} = .3/.59 = .051$$

Another Example (continued)

Interpretation: The probability that the transmission is faulty if the mechanic declares it good is only .051, quite small.

Generalizing Bayes Theorem

We can generalize Bayes Theorem to problems with more than 2 outcomes.

Suppose there are 3 nickel sized coins in a box.

Coin 1	Coin 2	Coin 3
2 headed	2 tailed	fair

Generalizing Bayes Theorem

You reach in and grab a coin at random and flip it without looking at the coin. It comes up heads. What is the probability that you have drawn the two headed coin (#1)?

$$\Pr(2H \mid \text{head}) = \frac{\Pr(2H)\Pr(\text{head} \mid 2H)}{\Pr(2H)\Pr(\text{head} \mid 2H) + \Pr(\text{fair})\Pr(\text{head} \mid \text{fair}) + \Pr(2T)\Pr(\text{head} \mid 2T)}$$

$$\Pr(2H \mid \text{head}) = \frac{(1/3)(1)}{(1/3)(1) + (1/3)(1/2) + (1/3)(0)} = (1/3)/(1/2) = 2/3 = .667$$

Part2: Naïve Bayes

- Concept
- example

Naïve Bayes Classifier

- What can we do if our data *d* has several attributes?
- <u>Naïve Bayes assumption:</u> Attributes that describe data instances are conditionally independent given the classification hypothesis

$$P(\mathbf{d} \mid h) = P(a_1, \ldots, a_T \mid h) = \prod_t P(a_t \mid h)$$

  - it is a simplifying assumption, obviously it may be violated in reality
  - in spite of that, it works well in practice
- The Bayesian classifier that uses the Naïve Bayes assumption and computes the Maximum A Posterio (MAP) hypothesis is called Naïve Bayes classifier
- One of the most practical learning methods
- Successful applications:
  - Medical Diagnosis
  - Text classification

Example. 'Play Tennis' data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Naïve Bayes solution

Classify any new datum instance $\mathbf{x}=(a_1,\ldots a_T)$ *as:*

$$h_{Naive\,Bayes} = \arg\max_h P(h)P(\mathbf{x}\mid h) = \arg\max_h P(h)\prod_t P(a_t \mid h)$$

- To do this based on training examples, we need to estimate the parameters from the training examples:
  - For each target value (hypothesis) *h*

    $$\hat{P}(h) := \text{estimate}\, P(h)$$

  - For each attribute value a_t of each datum instance

    $$\hat{P}(a_t \mid h) := \text{estimate}\, P(a_t \mid h)$$

Based on the examples in the table, classify the following datum ***x***:

x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \arg\max_{h\in[yes,no]} P(h)P(\mathbf{x}\,|\,h) = \arg\max_{h\in[yes,no]} P(h)\prod_t P(a_t\,|\,h)$$

$$= \arg\max_{h\in[yes,no]} P(h)P(Outlook = sunny\,|\,h)P(Temp = cool\,|\,h)P(Humidity = high\,|\,h)P(Wind = strong\,|\,h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong\,|\,PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong\,|\,PlayTennis = no) = 3/5 = 0.60$$

$$etc.$$

$$P(yes)P(sunny\,|\,yes)P(cool\,|\,yes)P(high\,|\,yes)P(strong\,|\,yes) = 0.0053$$

$$P(no)P(sunny\,|\,no)P(cool\,|\,no)P(high\,|\,no)P(strong\,|\,no) = \mathbf{0.0206}$$

$$\Rightarrow answer : PlayTennis(x) = no$$

Learning to classify text

- Learn from examples which articles are of interest
- The attributes are the words
- Observe the Naïve Bayes assumption just means that we have a random sequence model within each class!
- NB classifiers are one of the most effective for this task
- Resources for those interested:
  - Tom Mitchell: Machine Learning (book) Chapter 6.

NB vs. other classification methods

(a)

	NB	Rocchio	kNN	SVM
micro-avg-L (90 classes)	80	85	86	89
macro-avg (90 classes)	47	59	60	60

(b)

	NB	Rocchio	kNN	trees	SVM
earn	96	93	97	98	98
acq	88	65	92	90	94
money-fx	57	47	78	66	75
grain	79	68	82	85	95
crude	80	70	86	85	89
trade	64	65	77	73	76
interest	65	63	74	67	78
ship	85	49	79	74	86
wheat	70	69	77	93	92
corn	65	48	78	92	90
micro-avg (top 10)	82	65	82	88	92
micro-avg-D (118 classes)	75	62	n/a	n/a	87

Evaluation measure: F_1

Naive Bayes does pretty well, but some methods beat it consistently (e.g., SVM).

Remember

- Bayes' rule can be turned into a classifier
- Naive Bayes Classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attributes are independent given the class.
- Bayesian classification is a generative approach to classification

Resources

- Textbook reading (contains details about using Naïve Bayes for text classification):

  Tom Mitchell, Machine Learning (book), Chapter 6.

- Software: NB for classifying text:

  http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

- Useful reading for those interested to learn more about NB classification, beyond the scope of this module:

  http://www-2.cs.cmu.edu/~tom/NewChapters.html