Data Science Notes 10/7/24

Admin:

- Midterm due next Wednesday
- Code style is important
  - eg. Type hints, Model accuracy
- Lab 5 posted
  - Due Monday (Oct 21)

Informal Quiz:

1. P(A,B) “The probability A and B”
2. P(A,B) = P(B|A) P(A)
3. c. p(y | x)

Intro to Bayesian Models

- Helps us calculate probabilities using Bayes rule

Bayes Rule:

Bayes’ Theorem

- P(A,B) = P(A|B)P(B)
- P(A,B) = P(B|A)P(A)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Independence: P(A,B) = P(A)P(B) not true in general!!!

- Given data, use Bayes rule, calculate the probability of an event happening

$$\boxed{p(y=k|\boldsymbol{x}) = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{p(\boldsymbol{x})}}$$

- Components:
  - Evidence: p(x)
    - Data that we’re given / have already observed

- Prior: p(y = k)
    - Prior probability that we have in mind without seeing data
        - Eg. probability of anyone having the flu
- Posterior: p(y = k|x)
    - “Probability of y = k given x”
    - Probability of outcome *after* we’ve seen the evidence
- Likelihood: p(x | y = k)
    - “Probability of x given y = k”
    - Given an outcome, what is the prob of observing this set of features?

Examples

Spam mail

Bayesian Model for Trisomy 21

- C represents not down syndrome

Naive Bayes Algorithm

- “A Comparison of Event Models for Naive Bayes Text Classification” (5649 citations!)
- Text classification
- Goal: Classify documents into topics based on the words as features
- Eg: per document: 30% prob that its about sports, 50% that it’s about politics

- Single document $\vec{x} = [x_1, x_2, \ldots, x_p]^T$
- Multi-class response $y \in \{1, 2, \ldots, K\}$

- Goal: Classification $\hat{y} = argmax_{k=1,\ldots,K}\, p(y = k|\vec{x})$

Bayesian Model

$$\boxed{p(y = k|\vec{x}) = \frac{p(y = k)p(\vec{x}|y = k)}{p(\vec{x})}}$$

can ignore

Use the Bayesian Model to calculate probabilities for each K (class)

- Can ignore p(**x**)

1. Calculate the probability of the words, given the class y = k
    - All words are the features
    - Since this is a joint probability, apply Bayes rule
    - x1 = A, rest = B
    - Apply Bayes rule to continually break the join probability down until you run out of words
    - Multiply all probabilities together

Naive Bayes Assumption

- **Conditional Independence:** “feature j is independent from all other features given label k”

- Eg. Probability of something being a cat

Naïve Bayes Model

$$p(y = k|\vec{x}) \propto p(y = k) \prod_{j=1}^{p} p(x_j|y = k)$$

proportional to

- Given a document topic, all words are independent of each other
- To do classification, calculate this for all the k’s
    - Find the k with the highest probability

- Estimate based on training data
    - x vectors
    - class = documents
- N_k = # examples with label k
- What if N_k = 0?
    - Eg. training data doesn’t have any documents about Ballet
    - Theta is now 0, which makes everything 0 in the equation

Laplace Smoothing

- Technique to handle zero probability
- Theta is no longer 0

- Similarly, let $N_{k,j,v}$ = # examples with feature j = value v and class label k

- K is the number of different values that y can take

Example:
class y = tennis
    possible values: {yes, no}

N_{tennis} = yes = 7
$N_{yes,\ outlook,\ sunny}$ = 4
$Theta_{y,o,s}$ = (4 + 1) / (7 + 3) = 5 / 10 -> The value with laplace smoothing

* Using the Naive way, it would be 4/7

Handout 11