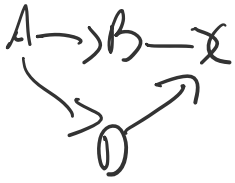
Bayes Nets Pt 2

Friday, March 23, 2018 8:58 AM

- Last time, What are Bayes Nets?
  - Graphical model: V = random variables (features)
    E = conditional Dependence relationships
    Conditional probability table for each variable



$\Rightarrow P(A,B,C,D) = P(A)\,P(B|A)\,P(D|A)\,P(C|B,D)$

Classification

- Class label (random variable), $c_j \in C$
  - c_j: specific class
  - C: random variable C
- Set of features $X = \langle x_1, x_2, \ldots, x_n \rangle$, evidence
- Max a posteriori hypothesis

$$c_{MAP} = \arg\max_j P(c_j \mid x_1, x_2, \ldots, x_n)$$

$P(c_j \mid x_1, x_2, \ldots, x_n)$: posterior distribution

$$= \arg\max_j \frac{P(x_1, x_2, \ldots, x_n \mid c_j)\,P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$P(x_1, x_2, \ldots, x_n)$: for normalization, so we can ignore it

$$= \arg\max_j P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

↑ Probability of evidence given class label

↑ Prior probability of observing c_j

- $P(c_j)$ is not always available
  - Assume uniform in that case
  - Maximum Likelihood estimate

$$c_{MLE} = P(x_1, x_2, x_3, \ldots, x_n \mid c_j)$$

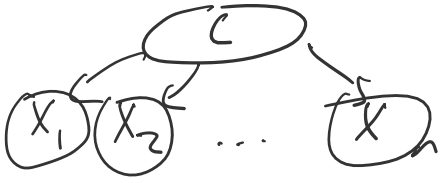
Typically, we stick with MAP hypothesis

- $c_{MAP} = P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$
- How hard is this to calculate in practice?
- $P(c_j)$ is easy!
  - Use frequencies of c_j to estimate $P(c_j)$

- $P(x_1, x_2, \ldots, x_n \mid c_j)$?. Super hard 😢
  - $O(|X|^n \cdot |C|)$ # of parameters
  - Superexponential growth in terms of data
  - Curse of dimensionality

- What do?

- Naive Bayes assumption: treat all features $X = \langle X_1, X_2, X_3, \ldots, X_n \rangle$ as independent given C



$P(X, Y|Z) = P(X|Z)\,P(Y|Z)$ if X, Y are conditionally independent given Z

Now!

$C_{MAP} = P(c_j) \prod_i P(x_i|c_j)$ ← usually much easier to calculate

- Do suffer information loss
- In practice, works well enough

Application: text classification

- input: some text, label {topic, or genre, or author}
- goal, use text as evidence to predict label

- Step 1, find text features
  - Word frequency
  - n-gram frequency: The quick brown fox
    - pros: temporal relation info — bigrams: The-quick, quick-brown, brown-fox
    - cons: data is sparse

- Position ≠ Brent way!!
  Don't be like Brent

- Each position is a variable

X_1 = The, X_2 = Quick, X_3 = Brown, X_4 = Fox

  - Doesn't scale
    - Unreasonable in practice

- Get rid of positional assumptions
  - Frequency of words are important!

- Bag of words model
  - Text classification depends solely on how often certain words are used