CS 581 – Advanced Artificial Intelligence

Topic: Naïve Bayes

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Bayes Classifier

- Input: $\vec{X} = \langle X_1, X_2, \dots, X_n \rangle$
- Output: Y
- Bayes classifier

$$P(Y \mid \vec{X}) = \frac{P(\vec{X} \mid Y)P(Y)}{P(\vec{X})} = \frac{P(Y)P(X_1, X_2, \dots, X_n \mid Y)}{P(X_1, X_2, \dots, X_n)}$$

$$P(X_1, X_2, \dots, X_n) = \sum_{y} P(Y = y)P(X_1, X_2, \dots, X_n \mid Y = y)$$

Assuming all variables are binary, how many independent parameters are needed for the Bayes classifier?

Excursion

- Maximum likelihood estimation
- Bayesian estimation

Bayes Classifier

- Assume a binary classification task, where the label Y is *spam* or ~*spam*
- Assume P(*spam*) = 0.4
- If you have seen an email X a times as *spam* and b times as ~*spam*
  - Using an MLE estimate for $P(X|Y)$, what is $P(Y|X)$?
  - Using an LS estimate for $P(X|Y)$, what is $P(Y|X)$?
  - What happens when either or both of a and b are zero?

Naïve Bayes Assumption

$$X_i \perp X_j \mid Y$$

Naïve Bayes

Bayes rule:

$$P(Y \mid X_1, X_2, \dots, X_n) = \frac{P(Y)P(X_1, X_2, \dots, X_n \mid Y)}{\sum_y P(y)P(X_1, X_2, \dots, X_n \mid y)}$$

Assuming $X_i \perp X_j \mid Y$,
naïve Bayes:

$$P(Y \mid X_1, X_2, \dots, X_n) = \frac{P(Y)\prod P(X_i \mid Y)}{\sum_y P(y)\prod P(X_i \mid y)}$$

Assuming all variables are binary, how many independent parameters are needed for the naive Bayes classifier?

Example

- See OneNote

Naïve Bayes Implementations

- Bernoulli / categorical naïve Bayes
  - Features are assumed to be binary / categorical
- Multinomial naïve Bayes
  - $P(\vec{X} \mid y)$ is a multinomial distribution
- Gaussian naïve Bayes
  - Each $p(x_i \mid y)$ is a Gaussian distribution

READING

- http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf
- https://en.wikipedia.org/wiki/Naive_Bayes_classifier
- https://scikit-learn.org/stable/modules/naive_bayes.html

Zero Probabilities

- Assume feature X_i is T for a particular object. Further,
- Assume $P(X_i = T|yes) = 0$ and $P(X_i = T|no) > 0$ and $P\left(X_j \mid yes\right) > 0$ and $P\left(X_j \mid no\right) > 0$ for all other features
  - What is $P\left(yes \mid \vec{X}\right)$?
- Assume $P(X_i = T|yes) = 0$ and $P(X_i = T|no) = 0$ and $P\left(X_j \mid yes\right) > 0$ and $P\left(X_j \mid no\right) > 0$ for all other features
  - What is $P\left(yes \mid \vec{X}\right)$?
- One solution: use LS for the parameter estimates

Multiplying Several Probability Numbers

- Assume we have 10,000 features
- What is $0.9^{10,000}$ using a computer?
- Try math.pow(0.9, 10000) in Python
- In Naïve Bayes,
  - $a = P(Y = T) \prod P(X_i | Y = T)$
  - $b = P(Y = F) \prod P(X_i | Y = F)$
  - $P\left(Y = T \middle| \vec{X}\right) = \frac{a}{a+b}$
  - If $a = b = 0$ in your code, then what?