Data Mining & Machine Learning

Yong Zheng
Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH
College of Computing

Schedule

- Review
- Classification by Naïve Bayes
- Classification Evaluation Metrics

Schedule

- Review
- Classification by Naïve Bayes
- Classification Evaluation Metrics

Review

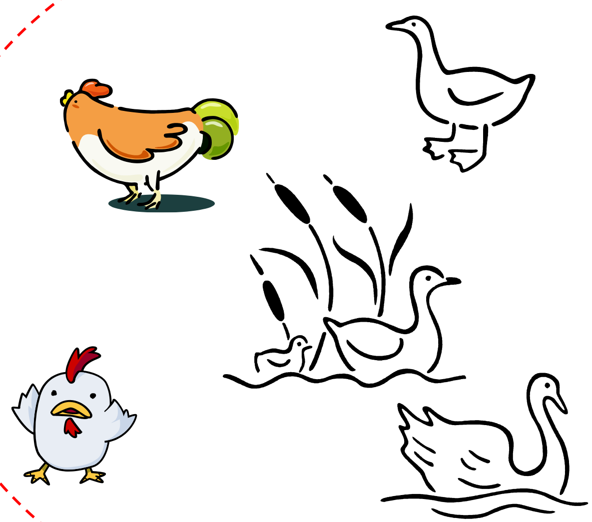
- Supervised vs Unsupervised Learning
- Supervised Learning: Classification
  - Standard Process
  - How to Split Data
  - Overfitting Issue
- Classification by KNN
  - How it works?
  - Three questions
    - Requirements on data, overfitting, learning process?

Schedule

- Review
- Classification by Naïve Bayes
- Classification Evaluation Metrics

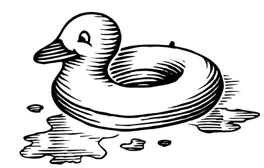
Naïve Bayes Classifier

- It is a probabilistic learning process
  - It is a simple classification algorithm too
  - You should have some preliminary knowledge about probability
  - There are some requirements to use the Naïve Bayes classifier
- Let’s see the duck example:



Training Records

Unseen Data, E



Pr (duck | E) = ?
Pr (chicken | E) = ?

Basic Concepts In Probability I

- P(*A* | *B*) is the probability of *A* given *B;* *conditional probability*

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

There are 10 examples here.

A: tiger = yes

B: color = orange

P (A) = 4/10 = 0.4

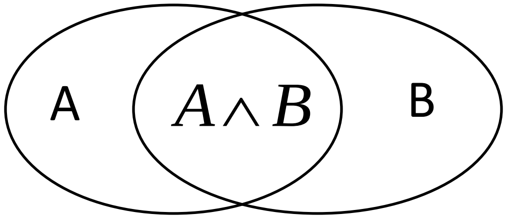
P (B) = 5/10 = 0.5

P (A | B) = ?

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
Orange	490	yes	yes
Orange	490	no	no
Orange	40	no	no
Orange	200	yes	no

Basic Concepts In Probability II

- P(*A* | *B*) is the probability of *A* given *B; conditional probability*
- Assumes that *B* is all and only information known.
- Defined by: $$\boxed{P(A \mid B) = \frac{P(A \wedge B)}{P(B)}}$$



- Bayes's Rule:
  Direct corollary of above definition

$$P(A \wedge B) = \frac{P(A \mid B)}{P(A)} = P(B \wedge A) = \frac{P(B \mid A)}{P(B)}$$

$$\Rightarrow P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

Naïve Bayes Classifier

- Let set of classes be $\{c_1, c_2, \ldots c_n\}$, e.g., c_1 = tiger, c_2 = lion
- Let E be description of an example (e.g., a vector with feature values)
- Determine class of E by computing for each class c_i

$$\boxed{P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}}$$

- P(E) can be determined since classes are complete and disjoint:

$$\boxed{\sum_{i=1}^{n} P(c_i \mid E) = \sum_{i=1}^{n} \frac{P(c_i)P(E \mid c_i)}{P(E)} = 1}$$

$$\boxed{P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)}$$

Naïve Bayes Classifier

- Determine class of *E* by computing for each class c_i

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$

$$P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)$$

- Note: E is a feature vector, instead of a single feature!!

  For example:

  c: tiger = yes

$$E = e_1 \wedge e_2 \wedge \cdots \wedge e_m$$

  E: color = orange, weight = 500 lbs, stripes = yes

- Assume features are independent given the class (c_i), *conditionally independent;* Therefore, we then only need to know $P(e_j \mid c_i)$ for each feature and category [IMPORTANT Assumption!!!]

$$P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$

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
- Generally, $P(X,Y|Z) = P(X|Z) \times P(Y|Z)$

Let's say you flip two regular coins:
A - Your first coin flip is heads
B - Your second coin flip is heads
C - Your first two flips were the same

What is the relationship between A and B?
How about [A and B] by given C?

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
- Generally, $P(X,Y|Z) = P(X|Z) \times P(Y|Z)$

There are a regular coin and a fake one (two heads)
I randomly choose one of them and toss it twice
A - Your first flip is heads
B - Your second flip is heads
C - Your select a regular coin

What is the relationship between A and B?
How about [A and B] by given C?

Naïve Bayes Classifier

- Determine class of *E* by computing for each class c_i

$$P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}$$

$$P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)$$

- Note: E is a feature vector, instead of a single feature!!

  For example:

  c: tiger = yes

  $$E = e_1 \wedge e_2 \wedge \cdots \wedge e_m$$

  E: color = orange, weight = 500 lbs, stripes = yes

- Assume features are independent given the class (c_i), *conditionally independent;* Therefore, we then only need to know $P(e_j \mid c_i)$ for each feature and category [IMPORTANT Assumption!!!]

$$P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$

Example: Naïve Bayes Classifier

- c1: tiger = yes; c2: tiger = no

  E: $\underbrace{\text{color = orange}}_{e1}$, $\underbrace{\text{weight >= 500 lbs}}_{e2}$, $\underbrace{\text{stripes = yes}}_{e3}$

Color	Weight (lbs)	Stripes	Tiger?
Orange	500	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

$$P(c1 \mid E) = \frac{P(c1)P(E \mid c1)}{P(E)}$$

$$P(E \mid c1) = \prod_{j=1}^{m} P(e_j \mid c1)$$

P(e1 | c1) = ¼ = 0.25
P(e2 | c1) = ¾ = 0.75
P(e3 | c1) = 4/4 = 1
P(E | c1) = 0.25*0.75*1 = 0.1875

P(e1 | c2) = 4/6 = 0.667
P(e2 | c2) = 1/6 = 0.167
P(e3 | c2) = 2/6= 0.333
P(E | c2) = 0.0371

Example: Naïve Bayes Classifier

- c1: tiger = yes; c2: tiger = no

  E: $\underbrace{\text{color = orange}}_{e1}$, $\underbrace{\text{weight = 500 lbs}}_{e2}$, $\underbrace{\text{stripes = yes}}_{e3}$

Color	Weight (lbs)	Stripes	Tiger?
Orange	500	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

$$P(c1 \mid E) = \frac{P(c1)P(E \mid c1)}{P(E)}$$

$$P(E \mid c1) = \prod_{j=1}^{m} P(e_j \mid c1\ )$$

P(E | c1) = 0.25*0.75*1 = 0.1875
P(E | c2) = 0.0371

$$P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)$$

P(c1) = 4/10 = 0.4
P(c2) = 6/10 = 0.6
P(E) = P(c1)P(E|c1) + P(c2)P(E|c2) = 0.09726
P(c1|E) = 0.4*0.1875/0.09726 = 0.7711

Example: Naïve Bayes Classifier

- c1: tiger = yes; c2: tiger = no

  E: $\underbrace{\text{color = orange}}_{e1}$, $\underbrace{\text{weight = 500}}_{e2}$ lbs, $\underbrace{\text{stripes = yes}}_{e3}$

Color	Weight (lbs)	Stripes	Tiger?
Orange	500	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

P(E | c1) = 0.25*0.75*1 = 0.1875
P(E | c2) = 0.0371
P(c1) = 4/10 = 0.4
P(c2) = 6/10 = 0.6
P(E) = P(c1)P(E|c1) + P(c2)P(E|c2) = 0.09726
P(c1|E) = 0.4*0.1875/0.09726 = 0.7711

$$P(c2|E) = \frac{P(c2)P(E|c2)}{P(E)}$$

P(c2|E) = 0.6*0.0371/0.09726 = 0.2289

Example: Naïve Bayes Classifier

- c1: tiger = yes; c2: tiger = no

  E: $\underbrace{\text{color = orange}}_{\text{e1}}$, $\underbrace{\text{weight = 500 lbs}}_{\text{e2}}$, $\underbrace{\text{stripes = yes}}_{\text{e3}}$

Color	Weight (lbs)	Stripes	Tiger?
Orange	500	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

P(c1|E) = 0.4*0.1875/0.09726 = 0.7711
P(c2|E) = 0.6*0.0371/0.09726 = 0.2289

P(c1|E) > P(c2|E)

We have more confidence to say we should trust c1

In other words, E should be classified as tiger!!!!

General Questions for Classifications

- What are the required data types by an algorithm
- Is there an overfitting problem?
- Is there a training-learning process?

Naïve Bayes: Categorical Only

Convert Numerical ones to categorical ones

- Numerical Features

Color	Weight (lbs)	Stripes	Tiger?
Orange	500	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
White	560	yes	yes

Weights = 500
Weights > 500

Create two groups

Naïve Bayes: Overfitting

- The overfitting issue is alleviated in Naïve Bayes, due to its nature – probabilistic model using priori probabilities
- But these probabilities may not be reliable, if we have limited knowledge on labeled data

Special Issue in Naïve Bayes

- Violation of Independence Assumption
  - It may be different to examine the assumptions
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the attribute value, i.e, $P(e_1 \mid c) = 0$
  - In this circumstance, $P(E \mid c)$ will be zero too during test

$$\boxed{P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)}$$

Relevant Issues

- Zero conditional probability Problem
  - For a remedy, conditional probabilities estimated with **Laplace smoothing**: **$P(e_i = A \mid C = c_j) = (n_c + m*p) / (n + m)$**
    - A is a value in the i-th feature; c_j is a value in the label
    - n_c = # of training instances for which **$e_i = A$ and $C = c_j$**
    - n = # of training instances for which $C = c_j$
    - m = a weight factor, usually m>=1 and could be big value, for example, the size of your training data
    - p = an estimate or a probability value to decrease m, usually, we can set p as 1/t and t is the number of unqiue values in e_i

Note: it is only applied to the $P(e_1 \mid c)$ component when it has the zero probability issue

Naïve Bayes: Summary

- Naïve Bayes is a probabilistic classification model
- It has one assumption: features are conditionally independent with labels
- Naïve Bayes Classification
  - Require categorical features
  - Overfitting is not serious
  - There is no learning process
  - Special issue: zero-probability issue
    Solution: Laplace smoothing

In-Class Practice

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

- Here, we have two classes C1="yes" (Positive) and C2="no" (Negative)
- Suppose we have new instance X = <sunny, mild, high, true>. How should it be classified?
- Compare Pr(P|X) and Pr(N|X)

Schedule

- Review
- Classification by Naïve Bayes
- Classification Evaluation Metrics

Classification Evaluation Metrics

- Accuracy is not the only metric
- Take binary classification for example

  Confusion Matrix

	Predicted Labels	
Actual Labels	+ (Yes)	- (No)
+ (Yes)	**True Positives (TP)**	**False Negatives (FN)**
- (No)	**False Positives (FP)**	**True Negatives (TN)**

Accuracy = (TP + TN)/All

Error rate = (FP + FN)/All

They are just overall metrics

It is possible that a model works well on overall, but very bad on a single label

Overall Acc = 90%, Acc on Positive label = 40%

Classification Evaluation Metrics

- Precision: exactness – what % of tuples that the classifier predicted as positive are positive

$$precision = \frac{TP}{TP + FP}$$

TP + FP = total number of predicted as positives

- Recall: completeness – what % of positive tuples did the classifier label as positive?

$$recall = \frac{TP}{TP + FN}$$

TP + FN = total number of actual positives

Classification Evaluation Metrics

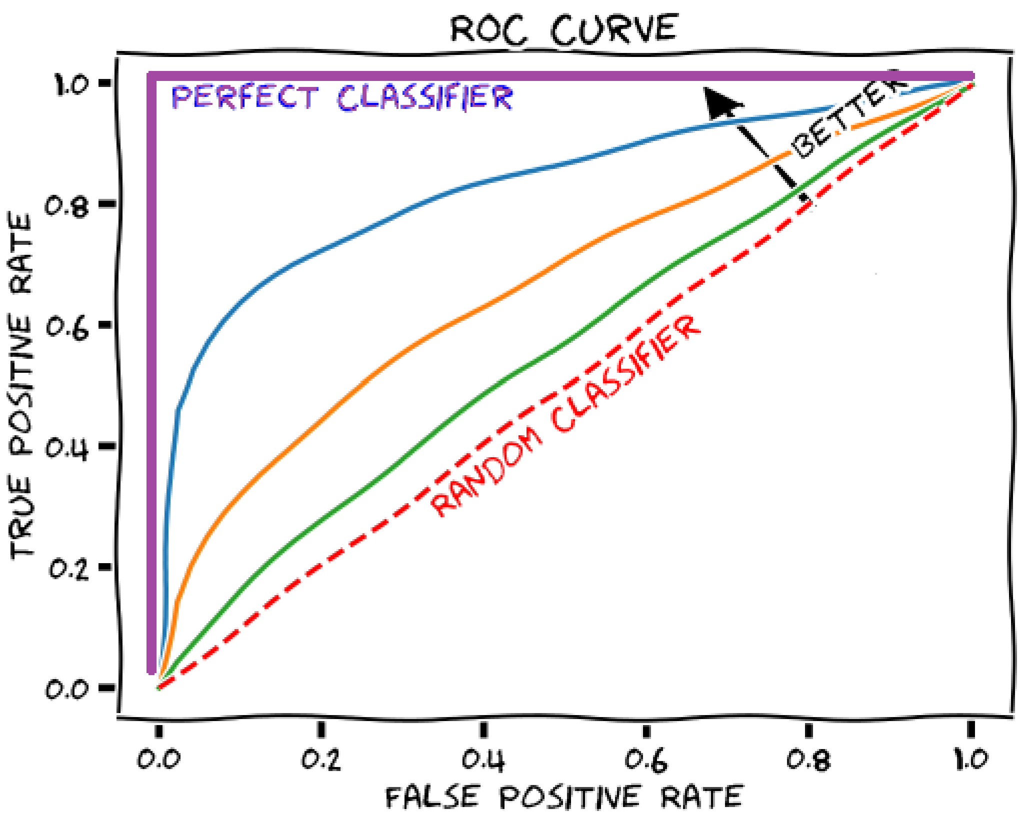
- F measure (F1 or F-score): harmonic mean of precision and recall

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- Sensitivity: True Positive recognition rate
  Sensitivity = TP/(TP+FN) = recall
- Specificity: True Negative recognition rate
  Specificity = TN/(FP+TN)

Classification Evaluation Metrics

- ROC Curve: false positive vs true positive rate
  false positive rate = 1 - specificity



You can observe the area under the curve. If this area is larger, a model is better.

Classification Evaluation Metrics

- In real-practice, reporting an overall metric is not enough. It is better to have the following combinations
  - Model 1: acc = 80%, acc_1 = 80%, acc_0 = 80%
  - Model 2: acc = 80%, acc_1 = 95%, acc_0 = 70%
- Overall metric + metric on positives & negatives
  - Acc/Err + Precision/Recall/F1 + Specificity
  - Acc/Err + ROC

In-Class Practice

- Calculate accuracy, err rate, precision, recall and F1, sensitivity and specificity for the following example

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	**90**	**210**	300
cancer = no	**140**	**9560**	9700
Total	230	9770	10000

Example: Metrics

- Answer

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	**90**	**210**	300	30.00 (*sensitivity*
cancer = no	**140**	**9560**	9700	98.56 (*specificity)*
Total	230	9770	10000	96.40 (*accuracy*)

Precision = 90/230 = 39.13%

Recall = 90/300 = 30.00%

In-Class Practice

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

- Here, we have two classes C1="yes" (Positive) and C2="no" (Negative)
- Suppose we have new instance X = <sunny, mild, high, true>. How should it be classified?
- Compare Pr(P|X) and Pr(N|X)