

Data Mining

- Classification Algorithms

-I

DKE - Gabriel Fung

Outline

- Introduction
- Classification Process
- Evaluating a classification model

P. 2

Introduction

P. 3

Classification

- A simple classification problem...
 - I know there are Salmon in this river.
 - When I pick up a fish from this river, can you tell me whether this fish is Salmon?
- Assume that you do not know how a Salmon looks like
 - Then... How to solve this problem?

P. 4

Classification

- Since you know nothing about Salmon and Tuna, the first thing you need to do is of course...LEARNING!
- Two types of learning
 1. Passive learning
 2. Active learning

P. 5

Different Kinds of Learning

- Passive learning
 - Find an expert.
 - The expert tells you all the characteristics of Salmon.
 - You simply memorize and apply what you have learned.
- Active learning
 - Find an expert.
 - The expert catches a lot of Fishes.
 - The experts only tells you which of them are Salmon, but does not tell you its characteristics.
 - You need to identify its characteristics by yourself by observing its features.

P. 6

Classification in Data Mining

- In data mining, we are always interested in active learning
 - You are an expert.
 - You catch a lot of Fishes.
 - You only tells the computer which of them are Salmon, but does not tell the computer its characteristics.
 - The computer identifies its characteristics by itself.
- Question:
 - As long as you are an expert, why don't you simply tell the characteristics of Salmon to the computer directly?

P. 7

Classification in Data Mining (cont'd)

- Answer:
 - Even an expert may sometimes find it difficult to generalize/extract/identify the characteristics of some observations...
- An example:
 - You have a lot of emails. You must know which of them are spam and which of them are not spam.
 - Yet, can you list ALL characteristics of the spam emails?
 - For active learning, you only need to tell the computer which of them are spam, and which of them are not.
 - The computer identifies their characteristics by itself by observing their differences.
 - So, you save lots of time in fact!

P. 8

Always Remember...

- From the data mining point of view...
 - Classification = Prediction = Forecasting
 - This is because the techniques are the same
- Classification is also known as "Supervised Learning"
 - There must be an "expert" (you) to "supervise" the computer.
 - In contrast, Clustering is known as "Unsupervised Learning". We will discuss it in the later lectures.

P. 9

Classification Process

P. 10

Terminologies

- Recall:
 - You catch a lot of Fishes.
 - You tells the computer which of them are Salmon.
 - The computer identifies their characteristics by itself.
- Terminologies:
 - Examples – The fishes that you have caught.
 - Class – Salmon and Not Salmon.
 - Positive examples – Fishes that belong to the class Salmon.
 - Negative examples – Fishes that do not belong to the class Salmon.
 - Model – What the computer has learned. The accuracy of the model depends on the learning algorithm.

P. 11

Learning and Operation

ID	Color	Size	...	Label
1	Pink	20cm	...	Salmon
2	Green	30cm	...	Not Salmon
...
N	Pink	18cm	...	Salmon

1. Archive Training Data

2. Choose an learning algorithm → **Model**

A new fish

→ **Model**

Yes (Salmon)

No (Not a Salmon)

■ Operation: P. 12

Binary-Class vs. Multi-Class

- Binary-Class Classification
 - Only two classes exists.
 - "Salmon" / "Not a Salmon"
 - "Cat" / "Dog"
 - "Sheep" / "Tiger"
- Multi-Class Classification
 - More than two classes.
 - "Salmon", "Tuna", "Shark", "Gold Fish"
 - Every multi-class classification problem can be solved by formulating a series of binary-class classification model.
 - How?

13

Binary-Class vs. Multi-Class (cont'd)

- Pay attention when formulating a model!!!
 - When classifying fishes...
 - If there are only two kinds of fish ("Salmon" and "Tuna")
 - It can be formulated as a simple binary-class classification problem
 - Trivial...
 - When classifying books...
 - If there are only two kinds of books ("Statistics" and "Algorithm")...
 - If a book can belong to either "Statistics" or "Algorithm" only...
 - A simple binary-class classification problem
 - If a book can belong to both "Statistics" and "Algorithm"...
 - Two binary-class classification problem

P. 14

Major Classification Algorithms

- In this course, we will discuss the following major learning algorithms:
 - Decision Tree
 - Nearest Neighbor
 - Naïve Bayes
 - Support Vector Machines

P. 15

Classifier Committee

- Also known as Ensemble Classifier.
- As the name implies, the decision is made by a set of classifiers.
- General idea
 - When a task involve expert's judgment, then the decision makes by N experts together is usually better than only one, if their decisions are properly combined.

P. 16

Two Simple Combination Techniques

- Majority Vote.
 - Simple voting! This strategy always performs surprisingly good!
 - Suppose there are 25 independent classifiers
 - Each classifier has error rate, $\epsilon = 0.35$
 - Probability that the ensemble classifier makes a wrong prediction:

$$\sum_{i=13}^{25} \binom{25}{i} \epsilon^i (1-\epsilon)^{25-i} = 0.06$$
- Weighted Linear Combination.
 - If a classifier is more reliable, then we value its decision higher.
 - We will discuss how to compute the reliability of a classifier shortly.
 - Usually performs even better than Majority Vote.

P. 17

Evaluation

P. 18

Testing

- After Learning and before using the model (operation), we need to test it first!
 - We need to "test" the model to see whether it "really learned" something.
 - To see how good (reliable) the model is.

P. 19

Testing

- Prepare the training data and testing data
 - Training Data and Testing Data will NEVER overlapped
 - Why?

ID	Color	Size	...	Label
1	Pink	20cm	...	Salmon
2	Green	30cm	...	Not Salmon
...
N	Pink	18cm	...	Salmon

Partition

ID	Color	Size	...	Label
1	Pink	20cm	...	Salmon
3	Green	32cm	...	Salmon
...
K	Black	24cm	...	Not Salmon

ID	Color	Size	...	Label
2	Green	30cm	...	Not Salmon
6	Grey	12cm	...	Not Salmon
...
M	Pink	18cm	...	Salmon

Training Data Testing Data

We will discuss how to partition shortly in the later slides

P. 20

Testing

Testing process:

ID	Color	...	Label
1	Pink	...	Salmon
2	Green	...	Not Salmon
...
N	Pink	...	Salmon

→

Model

↓

ID	Color	...	Label	Model's Decision
1	Pink	...	Salmon	Not Salmon
2	Green	...	Not Salmon	Salmon
...
N	Pink	...	Salmon	Salmon

This column is hidden to the model

This dataset MUST be different from the training data

Compare these two columns

Model Evaluation

- Metrics for Performance Evaluation
 - How to evaluate the performance of a model?
- Methods for Performance Evaluation
 - How to obtain reliable estimates?

Model Evaluation

- Metrics for Performance Evaluation
 - How to evaluate the performance of a model?
- Methods for Performance Evaluation
 - How to obtain reliable estimates?

Metrics for Performance Evaluation

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

A: TP (true positive) B: FN (false negative)

C: FP (false positive) D: TN (true negative)

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Accuracy (Note – This is NOT a good measurement):

Limitation of Accuracy

- Consider...
 - Total number of fishes in the testing examples = 10,000
 - Number of Non-Salmon = 9990
 - Number of Salmon = 10
- If model predicts everything to be class non-salmon, the accuracy is 9990/10000 = 99.9 %!!!
 - Accuracy is misleading because model cannot detect any Salmon.

Precision, Recall and F-Measurement

- Measuring the quality (effectiveness) of the model:
 - Precision, $p = \frac{A}{A+C}$
 - Recall, $r = \frac{A}{A+B}$
 - F-measure = $\frac{2rp}{r+p}$

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

P. 26

Evaluation of Classifiers: Receiver Operating Characteristics (ROC)

$tp\ rate \approx \frac{\text{Positives correctly classified}}{\text{Total positives}}$
 $fp\ rate \approx \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$

Point (0,0) issues no positive classifications.
 Point (1,1) issues positive classifications all the time.
 Point D (1,0) is perfect.
 Point (0,1) is the worst but can be reversed.
 Point C (on the line of $x=y$) indicates a random guess.
 So good classifiers appear at the upper left area of this graph.

A basic ROC graph showing five discrete classifiers.
http://home.comcast.net/~tom.fawcett/public_html/papers/
 27

Model Evaluation

- Metrics for Performance Evaluation
 - How to evaluate the performance of a model?
- Methods for Performance Evaluation
 - How to obtain reliable estimates? (i.e. how to partitioning the data?)

Methods of Estimation

- Holdout
 - Randomly take 70% of the examples as training and the remaining 30% as testing
 - Repeat the above procedure for several times (e.g. 10)

P. 29

Methods of Estimation

- Cross validation
 - Partition data into k disjoint subsets
 - Train on $(k-1)$ partitions, test on the remaining one
 - Repeat for all different combinations

P. 30



Methods of Estimation

- Leave-one-out estimation
 - Assume we have N examples.
 - Take $(N-1)$ examples as training, and the last one as testing
 - Repeat the experiment N times.

P. 31



Summary

P. 32



In this Lecture

- General procedure for constructing a classification model (classifier).
 - Partition the data into Training and Testing.
 - Train the classifier.
 - Test the classifier before using it.
 - Accuracy is not appropriate
 - Precision, Recall, F-Measure
- Combine the decisions of different classifiers
 - Majority vote
 - Linear Weight Combination
- Binary-class vs. Multi-class

P. 33



Reading

- Chapter 4: Classification, Introduction to Data Mining by P.N. Tan, M. Steinbach, and V. Kumar

P. 34