

CSCI 632 Notes

Clay L. McLeod

January 27, 2016

1 Machine Learning Overview

1.1 Supervised Learning

An **observation** is a d -dimensional vector X such that $X \in \mathbb{R}^d$.

The unknown nature of observation is called a **class**. We denote it by Y where $y \in \{1, 2, \dots, M\}$. For the purpose of this course, only discrete classes are considered (no regression).

The goal is to create a function $g(x) : \mathbb{R}^d \rightarrow \{1, \dots, M\}$ $g(x)$ one's guess of y given x . The classifier is $g(x)$. If $g(x) \neq y$.

Questions

1. How does one construct a good classifier?
2. How good can a classifier be?
3. Is classifier A better than classifier B ?
4. Can we estimate how good a classifier can be?
5. What is the best classifier?

The answer to all of these questions is yes: there are ways to find an upper bound on the performance of each algorithm and evaluate it empirically.

1.2 Unsupervised Learning

Same definition for an observation, except we don't have labels for the class in X . What approaches might this help us tackle?

Clustering

Unsupervised learning is directly related supervised learning. For instance: feature selection is probably the most important part of designing Machine Learning algorithms. Unsupervised learning helps us find good features for supervised learning algorithms.

Dimensionality reduction

As you increase the number of dimensions, you lose the ability to distinguish between two examples. Also, run time increases exponentially.

1.3 Semisupervised Learning

Partially labelled data where we try to gain some intuition. Usually involves a cost function instead of a solution set.

1.4 References

1. *A Probability Theory of Pattern Recognition* for Theoretical Design
2. *Machine Learning* for History of ML
3. *The Elements of Statistical Learning* for Statistical Vantagepoint
4. *Pattern Recognition and Machine Learning* (Textbook)
5. *Kernel Methods for Pattern Analysis* for Kernel Methods

2 Probability Review

In order to correctly analyze machine learning models and their correctness, we should first address some basic concepts in probability.

Definition: A probability space has 3 components.

1. A sample space, Ω , which is a set of all of the possible outcomes of a random process.
2. A family of sets, \mathfrak{S} representing the allowable events, where each set in \mathfrak{S} is a subset of Ω . \mathfrak{S} is a powerset of Ω .
3. A probability function $P_r : \mathfrak{S} \rightarrow R$ satisfying
 - (a) $\forall E \in \mathfrak{S}, 0 \leq P_r(E) \leq 1$
 - (b) $P_r(\Omega) = 1$
 - (c) $P_r(\bigcup_{i \geq 1} E_i) = \sum_{i \geq 1} P_r(E_i)$ if the RVs are independent.

Example: toss two dice

- $\Omega = \{(1, 1), (1, 2), \dots, (6, 6)\}$
- $\mathfrak{S} = \{\dots\} = |\mathfrak{S}| = 2^{36}$
- $P \rightarrow R$
 - $P((a, b)) = \frac{1}{36}, 1 \leq a, b \leq 6$
 - $P(E) = \sum_{(x,y) \in E} P((x, y)) = |E| \cdot \frac{1}{36}$

Lemma (Union bound)

Given: $\forall E_1, E_2 \subset \Omega$

Derived: $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \Rightarrow P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$

Lemma (Independence)

Given: \forall finite or countably infinite sequence of events E_1, E_2, \dots

Derived: $P_r(\bigcup_{i \geq 1} E_i) = \sum_{i \geq 1} P_r(E_i)$

Lemma (Inclusion-Exclusion principle)

Given: Let E_1, \dots, E_n be any of n events.

Derived: $P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i) - \sum_{i < j} P(E_i \cap E_j) + \sum_{i < j < k} P(E_i \cap E_j \cap E_k) \dots$

Definition

Two events E and F are independent if and only if

$$P(E \cap F) = P(E) \cdot P(F)$$

or, more generally the probability that *all* the events will happen is the same as the probability that *each* event will happened multiplied together.

Note: Independence \neq uncorrelated.

Definition

The conditional probability that the event E occurs given that event F occurs is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

or, written another way,

$$P(E \cap F) = P(E|F) \cdot P(F)$$

However,

$$P(E|F) = P(E)$$

when E and F are independent.

Theorem (Law of total probability)

Let E_1, \dots, E_n be mutually disjoint elements in Ω .

Theorem (Bayes' Law)

Assume that E_1, \dots, E_n are mutually disjoint sets such that

$$\bigcup_{i=1}^n E_i = \Omega$$

Then

$$P(E_j|B) = \frac{P(B|E_j) \cdot P(E_j)}{\sum_{i=1}^n P(B|E_i) \cdot P(E_i)}$$

This is proven by the combination of the law of conditional probability on the top and the law of total probability on the bottom.

Example

Two fair coins, biased coin($P(H) = \frac{2}{3}$). Assume that the output is HHT. What is the probability that the first coin was the biased coin?

- $B = \text{HHT}$
- $E_i = \text{ith coin toss is biased, } P(E_i) = \frac{1}{3}$.
- $P(E_1|B) = \frac{P(B|E_1) \cdot P(E_1)}{P(B)}$