Lecture Slides for

INTRODUCTION TO

Machine Learning
2nd Edition

ETHEM ALPAYDIN, modified by Leonardo Bobadilla
and some parts from
http://www.cs.tau.ac.il/~apartzin/MachineLearning/
© The MIT Press, 2010

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2m
CHAPTER 2:

Supervised Learning
Outline

• Previously:
  • Intro to Machine Learning
  • Applications
  • Logistics of the class

• This class: Supervised Learning (Sec 2.1-2.6)
  • Classification Learning a single class
  • Learning multiple classes
  • Theoretical aspects
  • Regression
Learning a Class from Examples

- Class C of a “family car”
  - **Prediction**: Is car $x$ a family car?
  - **Knowledge extraction**: What do people expect from a family car?
- **Output**: Positive (+) and negative (−) examples
- **Input representation**:
  - $x_1$: price, $x_2$: engine power
  - Expert suggestions
  - Ignore other attributes
Training set \( X \)

\[
X = \{x^t, r^t\}_{t=1}^N
\]

\[
r = \begin{cases} 
1 & \text{if } x \text{ is positive} \\
0 & \text{if } x \text{ is negative}
\end{cases}
\]

\[
x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}
\]
Class C
Class C

\[(p_1 \leq \text{price} \leq p_2) \ \text{AND} \ (e_1 \leq \text{engine power} \leq e_2)\]

How many parameters?

Assume class model (rectangle)
Hypothesis class $H$

$h(x) = \begin{cases} 1 & \text{if } h \text{ says } x \text{ is positive} \\ 0 & \text{if } h \text{ says } x \text{ is negative} \end{cases}$

Error of $h$ on $H$

$$E(h \mid X) = \sum_{t=1}^{N} 1(h(x^t) \neq r^t)$$

$\begin{array}{c}
\text{False positive} \\
\text{False negative}
\end{array}$
Generalization

- Problem of generalization: how well our hypothesis will correctly classify future examples

In our example: hypothesis is characterized by 4 numbers \((p1,p2,e1,e2)\)

Choose the best one
Include all positive and none negative
Infinitely many hypothesis for real-valued parameters
S, G, and the Version Space

The most specific hypothesis, $S$, and the most general hypothesis, $G$, form a version space $V_S$, which contains all consistent hypotheses $h \in H$ between $S$ and $G$.

(Mitchell, 1997)
Doubt

In some applications, a wrong decision is very costly

May reject an instance if fall between S (most specific) and G (most general)
Margin

- Choose $h$ with largest margin
Vapnik-Chervonenkis (VC) Dimension

Assumed that H (hypothesis space) includes true class C
H should be flexible enough or have enough capacity to include C
Need some measure of hypothesis space “flexibility” complexity
Can try to increase complexity of hypothesis space
VC Dimension

$N$ points can be labeled in $2^N$ ways as $+/-$

H shatters $N$ if there

exists $h \in H$ consistent

for any of these:

$\text{VC}(H) = N$

An axis-aligned rectangle

only!

An axis-aligned rectangle shatters 4 points only!
Probably Approximately Correct (PAC) Learning

Fix a probability of target classification error (planned future)

Actual error depends on training sample (past)

Want the actual probability error (actual future) be less than a target with high probability
Probably Approximately Correct (PAC) Learning

- How many training examples \( N \) should we have, such that with probability at least \( 1 - \delta \), \( h \) has error at most \( \varepsilon \)? (Blumer et al., 1989)

Let’s calculate how many samples we need for \( S \)

Each strip is at most \( \varepsilon/4 \)

Pr that we miss a strip \( 1 - \varepsilon/4 \)

Pr that \( N \) instances miss a strip \( (1 - \varepsilon/4)^N \)

Pr that \( N \) instances miss 4 strips \( 4(1 - \varepsilon/4)^N \)

\( 1 - 4(1 - \varepsilon/4)^N \geq 1 - \delta \) and \( (1 - x) \leq \exp(-x) \)

\( 4\exp(-\varepsilon N/4) \leq \delta \) and \( N \geq (4/\varepsilon)\log(4/\delta) \)
Probably Approximately Correct (PAC) Learning
Noise

Imprecision in recording the input attributes
Error in labeling data points (teacher noise)
Additional attributes not taken into account (hidden or latent)
Same price/engine with different label due to a color
Effect of this attributes modeled as a noise
Class boundary might be not simple
Need more complicated hypothesis space/model
Noise and Model Complexity

Use the simpler one because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance - Occam’s razor)
Occam's razor

If actual class is simple and there is mislabeling or noise, the simpler model will generalized better

Simpler model result in more errors on training set

Will generalized better, won’t try to explain noise in training sample

Simple explanations are more plausible!
Multiple Classes

General case K classes
Family, Sport, Luxury cars

Classes can overlap

Can use different/same hypothesis class

Fall into two classes? Sometimes worth to reject
Multiple Classes, $C_i$ for $i=1,\ldots,K$

Train hypotheses $h_i(x)$, $i = 1,\ldots,K$:

$$X = \{x^t, r^t\}_{t=1}^N$$

$$r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j, j \neq i \end{cases}$$

Lecture Notes for E Alpaydın 2010 Introduction to Machine Learning 2e © The MIT Press (V1.0)