Changelog: 14 October 2016, 30 October 2016

Decision Trees

Divide and Conquer, Loss Function, Inductive Bias

Lecture 3: Part 1

Marina Santini

Acknowledgements
Slides inspired by Daumé III (2015: 10-18)

Lecture 3: Required Reading

- Handout
- Daumé III (2015: 10-18)
- Witten et al. (2011: 99-108)

- Optional – Self Study: Witten et al. (2011: 192-203)

Outline

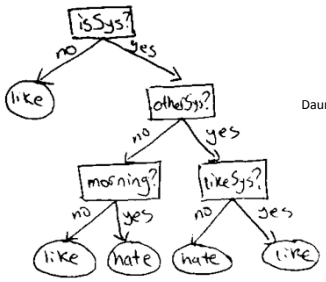
- Greedness
- Didive and Conquer
- Inductive Bias of the Decision Tree
- Loss Function
- Expected Loss
- Empirical Error
- Induction

Learning: Generalization-ability

- Predicting the future based on the past

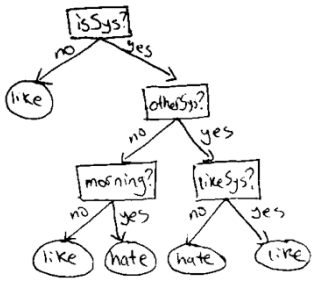
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*

Predict whether a student will like a course



Daumé III (2015)

Training Data



Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

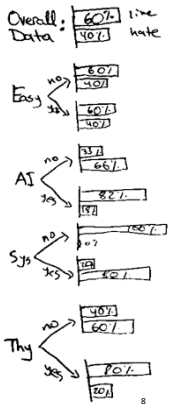
Problem Representation

- Questions = Features
- Answers = Feature Values
- Ratings = Class Labels

- An example is a set of feature values.
- Traning data is a set of examples associated with class labels.

"Greedy" Model: the most useful feature

- Histogram
- Root node



Divide & Conquer

- Divide:
  - Partition the data into 2 parts:
    - Yes part vs No part
- Conquer:
  - Recurse and run the Divide routine

The end of the cycle

… when it becomes useless to query on additional features

Decision Tree: Inductive Bias

The goal of the decision tree learning model is:
- to figure out what questions to ask
- in what order
- what answer to predict once you have asked enough questions

The inductive bias of decision trees:
- **The things that we want to learn to predict are more like the root node and less like the other branch nodes.**

Decision Tree: Informal Definition

A decision tree is:
- a flow-chart-like structure, where:
  - each internal (non-leaf) node denotes a test on an attribute
  - each branch represents the outcome of a test
  - each leaf (or terminal) node holds a class label

The topmost node in a tree is the root node.

Formalising the learning problem (1)

1. The loss function:

$$\textit{Classification: } \textbf{zero/one loss } \ell(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

Formalising the learning problem (2)

- Data Generating Distribution

$$D (x, y)$$

Expected Loss

1. The loss function
2. The data generationg distribution

$$\epsilon \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(y, f(\boldsymbol{x}))\big] = \sum_{(x,y)} \mathcal{D}(\boldsymbol{x}, y)\ell(y, f(\boldsymbol{x}))$$

How to read formulas: Expected Value

$$\epsilon \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(y, f(\boldsymbol{x}))\big]$$

ϵ = epsilon

$\triangleq$ =equal by definition to (or: is defined as)

$\mathbb{E}$ = blackboard-bold ***E***

(x,y) = sub the pair ***xy***

$\sim\mathcal{D}$ = overscript ***D***

$\ell(y, f(\boldsymbol{x}))$ = ***l*** of the pair ***y*** f of ***x***

$= \sum_{(x,y)} \mathcal{D}(\boldsymbol{x}, y)\ell(y, f(\boldsymbol{x}))$ = sum over all the pairs ***xy*** in and ***y*** times ***l*** of ***y*** and ***f*** of ***x***

Training Error

The training error is the average error over the training data

$$\hat{\epsilon} \triangleq \frac{1}{N}\sum_{n=1}^{N} \ell(y_n, f(\boldsymbol{x}_n))$$

How to read: the training error epsilon-hat is equal by definition to 1 over N of the Sum from n=1 to capital N of "l" of y and f of x.

Empirical Error (Alpaydin, 2014: 14)

The empirical error is the proportion of training instances where the predictions of h (the hypothesis = the informed guess) do not mach the required values given in X (the training set).

The error of the the hypothesis h given the training set X is:

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

Given:

- a loss function ***l*** and
- a sample ***d*** from some unknown distribution ***D***
- compute a function ***f*** that has low expected error ***ε*** over ***D*** with respect to ***l***.

Quiz 1: Training Error

How would you define a training error on a dataset:

1. Training error is the average loss over the training sample
2. Training error is the expected prediction error over an independent test sample
3. None of the above

Quiz 2: Distributions

$$\sum_{(x,y)} \mathcal{D}(x,y)\ell(y,f(x))$$

What kind of distribution is D in the formula above?

1. Normal
2. Unknown
3. None of the above

Quiz 3: Loss function

How would you define a loss function?

1. The loss function ***l(actual value, predicted value)*** characterizes how bad predictions are.
2. The loss function is an unknown distribution.
3. Both definitions are incorrect.

The end