Machine Learning 2007: Lecture 5

Instructor: Tim van Erven (Tim.van.Erven@cwi.nl)
Website: www.cwi.nl/~erven/teaching/0708/ml/

October 4, 2007

Overview

- **Organisational Matters**
- Probability Distributions and Random Variables
- Estimating Probabilities
- Information Theory
- The ‘Best’ Attribute in ID3
- Occam’s Razor

Course Organisation

- Don't work in pairs, unless explicitly allowed.
- Make sure your blackboard e-mailaddress works (I cannot change it) and that you read it.
- If you absolutely cannot attend the final exam, mail me.
- Exercise 2.1:

$$\mathcal{H} = \{\langle ?, ?, ?, ?, ?, ? \rangle, \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle, \langle \text{Warm}, ?, ?, ?, ?, ? \rangle, \ldots, \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\}$$

should be

$$\mathcal{H} = \{\langle ?, ?, ?, ?, ?, ? \rangle, \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle, \langle \text{Cloudy}, ?, ?, ?, ?, ? \rangle, \ldots, \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle\}$$

This Lecture versus Mitchell

Mitchell:

- Read: Chapter 3 of Mitchell.

This Lecture:

- More background on probability distributions and random variables.
- More about information theory than in Mitchell.

Overview

- Organisational Matters
- **Probability Distributions and Random Variables**
- Estimating Probabilities
- Information Theory
- The ‘Best’ Attribute in ID3
- Occam’s Razor

Probability Distributions Reminder

Given **sample space** $\Omega = \{\omega_1, \ldots, \omega_k\}$ a **probability mass function** $p(\omega_i)$ is a function that assigns a weight to each outcome ω_i such that

- $0 \leq p(\omega_i) \leq 1$
- $p(\omega_1) + \ldots + p(\omega_k) = 1$.

This mass function uniquely defines a **probability distribution** $P(\mathcal{E})$ that assigns probability

$$P(\mathcal{E}) = \sum_{\{i|\omega_i \in \mathcal{E}\}} p(\omega_i)$$

to any event $\mathcal{E} \subseteq \Omega$.

Conditional Distributions

Getting New Information:

- Let P be a probability distribution on sample space Ω.
- Suppose we are given the information that we will get an outcome in $\mathcal{E}_2 \subseteq \Omega$.
- How should we update P to take this into account?

Conditional Distributions

Getting New Information:

- Let P be a probability distribution on sample space Ω.
- Suppose we are given the information that we will get an outcome in $\mathcal{E}_2 \subseteq \Omega$.
- How should we update P to take this into account?

The Conditional Distribution:

- Make a new **conditional distribution** $P(\mathcal{E}_1 \mid \mathcal{E}_2)$ on Ω.
- The **conditional probability** of event $\mathcal{E}_1 \subseteq \Omega$ is:

$$P(\mathcal{E}_1 \mid \mathcal{E}_2) = \frac{P(\mathcal{E}_1 \cap \mathcal{E}_2)}{P(\mathcal{E}_2)},$$

(assuming $P(\mathcal{E}_2) > 0$).

Random Variables

Given sample space $\Omega = \{\omega_1, \ldots, \omega_k\}$, a **random variable** X assigns a number $X(\omega)$ to each outcome $\omega \in \Omega$: It is a function from Ω to $\mathbb{R}$.

Example:

Suppose $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ describes the possible outcomes of two coin flips (H = heads; T = tails). Then we might define a random variable that counts the number of heads:

ω	$X(\omega)$
HH	2
HT	1
TH	1
TT	0

Overview

- Organisational Matters
- Probability Distributions and Random Variables
- **Estimating Probabilities**
- Information Theory
- The ‘Best’ Attribute in ID3
- Occam’s Razor

Probabilistic Data

A Loaded Die:

- We roll a die n times and get data $D = y_1, \ldots, y_n$.
- For example $D = 6, 2, 6, 6, 6, 3, 6$.
- We consider it possible that the die has been loaded: Some sides may have been made heavier than others.
- How do we describe the statistical regularity in our data?

Probabilistic Data

A Loaded Die:

- We roll a die n times and get data $D = y_1, \ldots, y_n$.
- For example $D = 6, 2, 6, 6, 6, 3, 6$.
- We consider it possible that the die has been loaded: Some sides may have been made heavier than others.
- How do we describe the statistical regularity in our data?

Describing the Die Using a Distribution:

- View each throw as an outcome y from sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The probability distribution P of y depends on the die.
- For example, if the die has not been loaded, then P assigns the same probability $1/6$ to all outcomes.

Estimating Probabilities

Motivation:

- Suppose we get data $D = y_1, \ldots, y_n$, where each y_i has the same probability distribution P.
- We want to predict $P(y_{n+1} = 6)$.
- But we don't know P! We only see the data.

Estimating Probabilities

Motivation:

- Suppose we get data $D = y_1, \ldots, y_n$, where each y_i has the same probability distribution P.
- We want to predict $P(y_{n+1} = 6)$.
- But we don't know P! We only see the data.

Estimating the Probability of an Event:

- Then if we have a lot of data (n is large), we can estimate the probability P of any event $\mathcal{E}$ by the relative frequency of the occurrence of the event in D.
- For example, suppose $D = 6, 2, 6, 6, 6, 3, 6$. Then our estimate of $P(y = 6)$ will be $5/7$.

Overview

- Organisational Matters
- Probability Distributions and Random Variables
- Estimating Probabilities
- **Information Theory**
- The ‘Best’ Attribute in ID3
- Occam’s Razor

Information Theory

Set-up: Alice sends information to Bob over a (possibly noisy) communication channel, for example a telegraph line.

Information Theory

Set-up: Alice sends information to Bob over a (possibly noisy) communication channel, for example a telegraph line.

Important Concepts (informally):

- Entropy $H(X)$ of random variable X: minimum expected number of binary questions needed to determine $X(\omega)$.
- Mutual information $I(X;Y)$ of X and Y: How much information do we get about $X(\omega)$ by being told $Y(\omega)$?

Information Theory

Set-up: Alice sends information to Bob over a (possibly noisy) communication channel, for example a telegraph line.

Important Concepts (informally):

- Entropy $H(X)$ of random variable X: minimum expected number of binary questions needed to determine $X(\omega)$.
- Mutual information $I(X;Y)$ of X and Y: How much information do we get about $X(\omega)$ by being told $Y(\omega)$?

History:

- Until the early 1940s people thought that increasing the transmission rate of information over a communication channel increases the probability of error.
- Then **C.E. Shannon** showed that this is not true as long as the communication rate is below the channel capacity C, which is defined using mutual information.

Entropy

Definition:

The entropy $H(X)$ of a random variable X is defined as

$$H(X) = \sum_{x} P(X = x) \cdot (-\log_2 P(X = x)),$$

where x ranges over the possible values of $X(\omega)$.

Entropy

Definition:

The entropy $H(X)$ of a random variable X is defined as

$$H(X) = \sum_x P(X = x) \cdot (-\log_2 P(X = x)),$$

where x ranges over the possible values of $X(\omega)$.

Remarks:

- Entropy can be interpreted as the minimum expected number of binary questions needed to determine $X(\omega)$.
- Hence it measures our uncertainty about $X(\omega)$.

Entropy

Definition:

The entropy $H(X)$ of a random variable X is defined as

$$H(X) = \sum_x P(X = x) \cdot (-\log_2 P(X = x)),$$

where x ranges over the possible values of $X(\omega)$.

Remarks:

- Entropy can be interpreted as the minimum expected number of binary questions needed to determine $X(\omega)$.
- Hence it measures our uncertainty about $X(\omega)$.
- Note that if $P(X = x) = 0$, then $P(X = x) \cdot (-\log_2 P(X = x)) = 0 \log_2 0$ is undefined. We therefore define $0 \log_2 0 = 0$.
- Mitchell uses estimated values for $P(X = x)$.

Entropy Example

- Suppose

x	$P(X = x)$
0	1/4
1	1/2
2	1/4
3	0

- Then

$$\begin{aligned} H(X) = {} & P(X=0) \cdot -\log_2 P(X=0) \\ & + P(X=1) \cdot -\log_2 P(X=1) \\ & + P(X=2) \cdot -\log_2 P(X=2) \\ & + P(X=3) \cdot -\log_2 P(X=3) \\ = {} & 1/4 \cdot 2 + 1/2 \cdot 1 + 1/4 \cdot 2 + 0 \log_2 0 \\ = {} & 1.5 \end{aligned}$$

Conditional Entropy

Suppose X and Y are random variables.

Known $Y(\omega)$:
Suppose we have been told that $Y(\omega) = y$. Then we should use the conditional distribution $P(X \mid Y(\omega) = y)$ to compute the entropy of X:

$$H(X|Y = y) = \sum_{x} P(X = x|Y = y) \cdot (-\log_2 P(X = x|Y = y)).$$

Definition of Conditional Entropy:

The conditional entropy $H(X|Y)$ of X given Y is defined as

$$H(X|Y) = \sum_{y} P(Y = y) H(X|Y = y),$$

where y ranges over the possible values of $Y(\omega)$.

Mutual Information

Definition:

The mutual information $I(X;Y)$ between random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X \mid Y)$$

Remarks:

- $I(X;Y)$ may be interpreted as the expected reduction in our uncertainty about $X(\omega)$ by hearing the value of $Y(\omega)$.
- This is the amount of information we get about the value of $X(\omega)$ by being told the value of $Y(\omega)$.

Mutual Information Example

Suppose $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ and P assigns the same probability $(1/4)$ to all outcomes. Let X count the number of heads and Y indicate whether the first and the second outcome are the same:

ω	$X(\omega)$	$Y(\omega)$
HH	2	0
HT	1	1
TH	1	1
TT	0	0

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= 1.5 - P(Y=0)H(X|Y=0) \\ &\quad - P(Y=1)H(X|Y=1) \\ &= 1.5 - (1/2 \cdot 1 + 1/2 \cdot 0) = 1 \end{aligned}$$

Overview

- Organisational Matters
- Probability Distributions and Random Variables
- Estimating Probabilities
- Information Theory
- **The ‘Best’ Attribute in ID3**
- Occam’s Razor

The ID3 *Algorithm Reminder*

General:

- Learns a decision tree from data.
- Hence does classification.

Main Ideas:

1. Start by selecting a root attribute for the tree.
2. Then grow the tree by adding more and more attributes to it.
3. Stop growing the tree when it is consistent with all the data.

The ID3 *Algorithm*

D = data; $D_{a,v}$ = data such that $\mathbf{x}$ has value v for attribute x_a;
A = set of available features/attributes

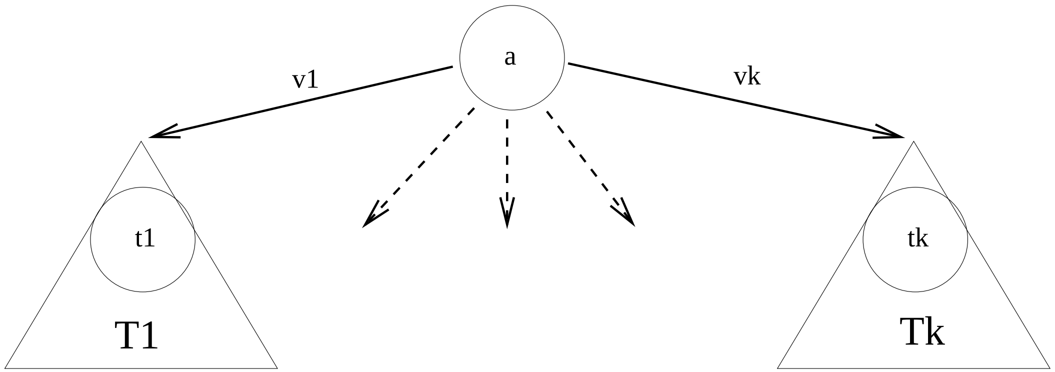
ID3(D, A)

z = the most common label y in D
if y is the same for all examples in D or $A = \emptyset$ **then**
  return $T = (\{z\}, \emptyset)$

Select the ‘best’ attribute $a \in A$ with values $v_1, \ldots, v_k$.

$$T_i = \begin{cases} (\{z\}, \emptyset) & \text{if } D_{a,v_i} = \emptyset \\ \text{ID3}(D_{a,v_i}, A \setminus \{a\}) & \text{otherwise} \end{cases}$$

return



An Attribute is a Random Variable

- In classification an outcome is $\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} \in \Omega = \mathcal{X} \times \mathcal{Y}$.
- For each attribute a, we define a random variable X_a that gives the value of the attribute:

$$X_a \left(\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} \right) = x_a.$$

- Likewise, we define a random variable Y that gives the value of the label:

$$Y \left(\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} \right) = y.$$

The ‘Best’ Attribute

The ‘Best’ Attribute:

ID3 selects the attribute a that gives the most information about the label:

$$\max_a I(Y; X_a)$$

The ‘Best’ Attribute

The ‘Best’ Attribute:

ID3 selects the attribute a that gives the most information about the label:

$$\max_a I(Y;X_a)$$

It Has to Estimate Probabilities:

To compute $I(Y;X_a)$, ID3 has to estimate $P(Y=y)$, $P(X_a=v)$, and $P(Y=y \mid X_a=v)$ for all possible labels y and values v of attribute a.

Remarks:

- Mitchell calls the mutual information with estimated probabilities the **information gain**.

A Second Discussion of ID3

The Inductive Bias of ID3:

- Smaller decision trees are preferred over bigger decision trees.
- Trees that place attributes that give the most information about the labels close to the root are preferred over trees that do not.
- (When) does a preference for shorter trees make sense?

Overview

- Organisational Matters
- Probability Distributions and Random Variables
- Estimating Probabilities
- Information Theory
- The ‘Best’ Attribute in ID3
- **Occam’s Razor**

Occam’s Razor

Originally:

The fourteenth century logician and natural philosopher William of Ockham stated:

> “What can be explained with fewer things is vainly explained with more.”

Remarks:

- This **inductive bias** is applied informally throughout the sciences: physicists prefer simpler explanations for the motions of the planets over more complex explanations.
- As Mitchell puts it: Prefer the simplest hypothesis (e.g the one with the smallest decision tree) that fits the data.
- ID3 follows Occam’s razor if we think that smaller decision trees are simpler than bigger decision trees.

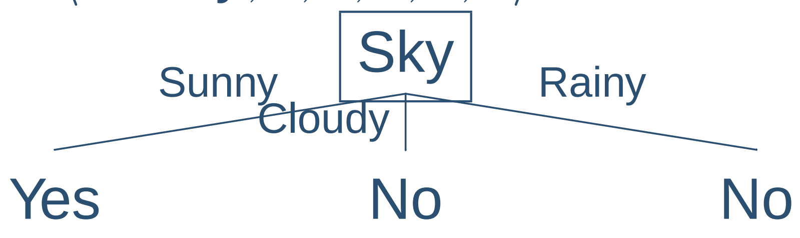
Does Occam's Razor Make Sense?

A Motivation of Occam's Razor:

- There are fewer simple hypotheses than complex hypotheses (e.g. fewer small decision trees than big decision trees)
- It is therefore less likely to be a coincidence when a simple hypothesis fits the training data well.

Dependence on the Language for Hypotheses:

- The same hypothesis in the EnjoySport example can be represented in different ways:
  - A list of constraints: $\langle$Sunny$, ?, ?, ?, ?, ?\rangle$
  - A decision tree:



- What appears simpler in one representation may look more complex in another, and vice versa.

Conclusions

Doubts:

- Occam’s razor depends on the language we use to describe hypotheses.
- Without knowing the language, Occam’s razor is too imprecise: What is simple?

Conclusions

Doubts:

- Occam’s razor depends on the language we use to describe hypotheses.
- Without knowing the language, Occam’s razor is too imprecise: What is simple?

Encouraging Thoughts:

- Occam’s razor makes sense if our language for describing hypotheses is such that simpler hypotheses are better than more complex hypotheses.
- Hence if we accept Occam’s razor, then we still have to specify our inductive bias by choosing a language for hypotheses.

Conclusions

Doubts:

- Occam's razor depends on the language we use to describe hypotheses.
- Without knowing the language, Occam's razor is too imprecise: What is simple?

Encouraging Thoughts:

- Occam's razor makes sense if our language for describing hypotheses is such that simpler hypotheses are better than more complex hypotheses.
- Hence if we accept Occam's razor, then we still have to specify our inductive bias by choosing a language for hypotheses.
- Maybe that is not such a bad way to specify inductive bias.
- This idea is formalised by the minimum description length principle, which turns out to have many elegant properties.

Overview

- Organisational Matters
- Probability Distributions and Random Variables
- Estimating Probabilities
- Information Theory
- The ‘Best’ Attribute in ID3
- Occam’s Razor

References

- T.M. Cover and J.A. Thomas, "Elements of Information Theory," 1991
- A.N. Shiryaev, "Probability", Second Edition, 1996
- A. Barron, "Logically Smooth Density Estimation", PhD-thesis, 1985
- P. Grünwald, "The Minimum Description Length Principle", 2007