Machine Learning for Intelligent Systems

Lecture 4: Prediction and Overfitting

Reading: UML 2.1-2.2, 18.2

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

Inductive Learning

Instance Space:

Instance space X including feature representation.

Target Attributes (Labels):

A set Y of labels.

Hidden target function:

An unknown function $f: X \rightarrow Y$ that is how instance are labeled in life.

Training Data:

A set of labeled pairs $\left(x, f(x)\right) \in X \times Y$ that we have seen before.

Hypothesis space: Last Lecture

A predetermined set H of functions in which we look for $h: X \rightarrow Y$.

Inductive Learning

Given a large enough number of training examples,

and given a hypothesis space H,

learn a hypothesis $h \in H$ that approximates $f(\cdot)$

Inductive Learning

Instance Space:

Instance space X including feature representation.

Target Attributes (Labels):

A set Y of labels.

Hidden target function:

An unknown function $f: X \rightarrow Y$ that is how instance are labeled in life.

Training Data: This Lecture

A set S of labeled pairs $\big(x, f(x)\big) \in X \times Y$ that we have seen before.

Hypothesis space:

A predetermined set H of functions in which we look for $h: X \rightarrow Y$.

Inductive Learning

Given a large enough number of training examples,

and given a hypothesis space H,

learn a hypothesis $h \in H$ that approximates $f(\cdot)$

How does performance of h on S translates to unseen instances.

World as a Distribution

A particular **instance of a learning problem** can be described as a joint probability distribution $P(X, Y)$ over $X \times Y$.

For example:

- A^+ Homework: $P(X, Y)$ indicates the probability that a homework with features X will receive A^+ label Y.

$$P(X = (\overset{\text{Correct?}}{\text{Complete}},\ \overset{\text{Colored?}}{\text{Yes}},\ \overset{\text{Original?}}{\text{Yes}},\ \overset{\text{Presentation?}}{\text{Clear}},\ \overset{\text{Latex?}}{\text{No}}),\ \overset{A^+?}{\text{Yes}})$$

- Tasty Apple: $P(X, Y)$ indicates the probability that an apple with features X has tastiness label Y.

$$P(X = (\overset{\text{Farm?}}{\text{A}},\ \overset{\text{Color?}}{\text{Red}},\ \overset{\text{Size?}}{\text{Medium}},\ \overset{\text{Firmness?}}{\text{Crunchy}}),\ \overset{\text{Tasty?}}{\text{Yes}})$$

Learning as Prediction

Chain rule: Sampling as a two step procedure

$$\mathrm{P}(\mathrm{X}, \mathrm{Y}) \quad = \quad \mathrm{P}(\mathrm{X}) \quad \times \quad \mathrm{P}(\mathrm{Y}|\mathrm{X})$$

$\mathrm{P}(\mathrm{X})$: Prob. the world produces instance with representation X

Example 1: X is homework representation:

$x_1 = (\text{complete}, \text{Yes}, \text{Yes}, \text{Clear}, \text{No}), x_2 = (\text{guessing}, \text{Yes}, \text{Yes}, \text{Clear}, \text{Yes})$

With prob. $\mathrm{P}(\mathrm{X} = x_1) = 0.2$ and $\mathrm{P}(\mathrm{X} = x_2) = 0.0001$.

Example 2: X is an apple representation:

$x_1 = (\text{A}, \text{red}, \ \text{medium}, \ \text{crunchy}), x_2 = (\text{B}, \text{green}, \text{small}, \text{soft})$

With prob. $\mathrm{P}(\mathrm{X} = x_1) = 0.25$ and $\mathrm{P}(\mathrm{X} = x_2) = 0.01$.

$\mathrm{P}(\mathrm{Y}|\mathrm{X})$: Prob. of seeing label Y on instance X.

Example 1: Prob. the teacher assign A^{+} to homework X. Deterministic label.

$\mathrm{P}(\mathrm{Y} = yes|x_1) = 1$ and $\mathrm{P}(\mathrm{Y} = yes|x_2) = 0$.

Example 2: Prob. an apple with features X is tasty. Non-deterministic label.

$\mathrm{P}(\mathrm{Y} = yes|x_1) = 0.9$ and $\mathrm{P}(\mathrm{Y} = yes|x_2) = 0.1$.

How is the data generated?

Independently: Seeing a labeled instance doesn’t affect prob. of others.

→ Y_i depends on X_i, but NOT on X_j and Y_j, for $i \neq j$.

→ What does it mean for homeworks? Cheating?

→ For apples? A disease affecting many apple trees?

Identically: $P(X)$ and $P(Y|X)$ don’t change over time.

→ $P(X_i = x, Y_i = y) = P\left(X_j = x, Y_j = y\right)$ for all i and j.

→ Quality of students changes over time? The selection criterion?

→ What about for apples?

Independently Identically Distributed (i.i.d)

A sample $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ is **independently identically distributed** according to $P(X, Y)$ if

$$\Pr(S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}) = \prod_{i=1}^{m} P(X = x_i, Y = y_i)$$

Sample & Generalization Errors

$\Delta(a, b)$ is the 0/1-loss function. i.e.,

$$\Delta(a, b) = \begin{cases} 0 & if\ (a = b) \\ 1 & otherwise \end{cases}$$

Sample (Empirical) Error

Sample error of hypothesis h on samples $\mathrm{S} = \{(x_1, y_1),\ \ldots,\ (x_m, y_m)\}$, denoted by $err_S(h)$ is

$$err_S(h) = \frac{1}{m}\sum_{i=1}^{m}\Delta(h(x_i), y_i)$$

Generalization (Prediction/true) Error

Generalization error of hypothesis h on distribution $P(X, Y)$, denoted by $err_P(h)$ is

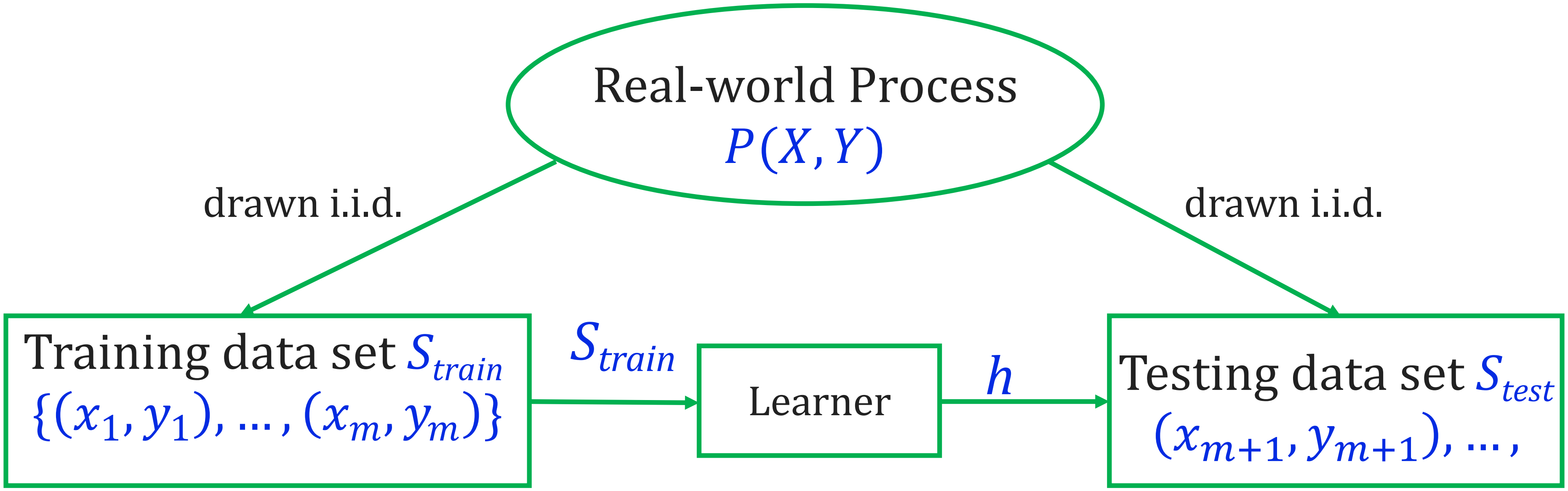
$$err_P(h) = \mathbb{E}_{(x,y)\sim P}[\Delta(h(x), y)]$$

Prediction as Learning

Goal: Find h with small prediction error $err_P(h)$ on $P(X,Y)$.

Strategy: Find an h with small sample error $err_{S_{train}}(h)$ on training dataset S_{train}.

Test the learned h to measure its **test error** $err_{S_{test}}(h)$ on a separate testing data set S_{test}.


Real-world Process
$P(X,Y)$
drawn i.i.d.
drawn i.i.d.
Training data set S_{train}
$\{(x_1,y_1),\dots,(x_m,y_m)\}$
S_{train}
Learner
h
Testing data set S_{test}
$(x_{m+1},y_{m+1}),\dots,$

Example: Text Classification

- Task: Learn rule that classifies Reuters Business News
  - Class +: “Corporate Acquisitions”
  - Class -: Other articles
  - 2000 training instances
- Representation:
  - Boolean attributes, indicating presence of a keyword in article
  - 9947 such keywords (more accurately, word “stems”)

LAROCHE STARTS BID FOR NECO SHARES +

Investor David F. La Roche of North Kingstown, R.I., said he is offering to purchase 170,000 common shares of NECO Enterprises Inc at 26 dlrs each. He said the successful completion of the offer, plus shares he already owns, would give him 50.5 pct of NECO's 962,016 common shares. La Roche said he may buy more, and possible all NECO shares. He said the offer and withdrawal rights will expire at 1630 EST/2130 gmt, March 30, 1987.

SALANT CORP 1ST QTR
FEB 28 NET −

Oper shr profit seven cts vs loss 12 cts. Oper net profit 216,000 vs loss 401,000. Sales 21.4 mln vs 24.9 mln. NOTE: Current year net excludes 142,000 dlr tax credit. Company operating in Chapter 11 bankruptcy.

Decision Tree for “Corporate Acq.”

```
• vs = 1: -
• vs = 0:
• |   export = 1:
…
• |   export = 0:
• |   |   rate = 1:
• |   |   |   stake = 1: +
• |   |   |   stake = 0:
• |   |   |   |   debenture = 1: +
• |   |   |   |   debenture = 0:
• |   |   |   |   |   takeover = 1: +
• |   |   |   |   |   takeover = 0:
• |   |   |   |   |   |   file = 0: -
• |   |   |   |   |   |   file = 1:
• |   |   |   |   |   |   |   share = 1: +
• |   |   |   |   |   |   |   share = 0: -
… and many more
```

Learned tree:

- has 437 nodes
- is consistent
- $err_{S_{train}}(h) = 0$

Accuracy of learned tree:

- $err_{S_{test}}(h) = 0.11$

Note: word stems expanded for improved readability.

Overfitting

Overfitting

Hypothesis h overfits to the training data S if $err_P(h) \gg err_S(h)$.

The issue with overfitting it that there could have been another hypothesis h', such $err_S(h) \lesssim err_S(h')$ but $err_P(h) \gg err_P(h')$.

Question: Does h_S overfit on samples $S = \{(x_1, f(x_1)), \dots\}$?

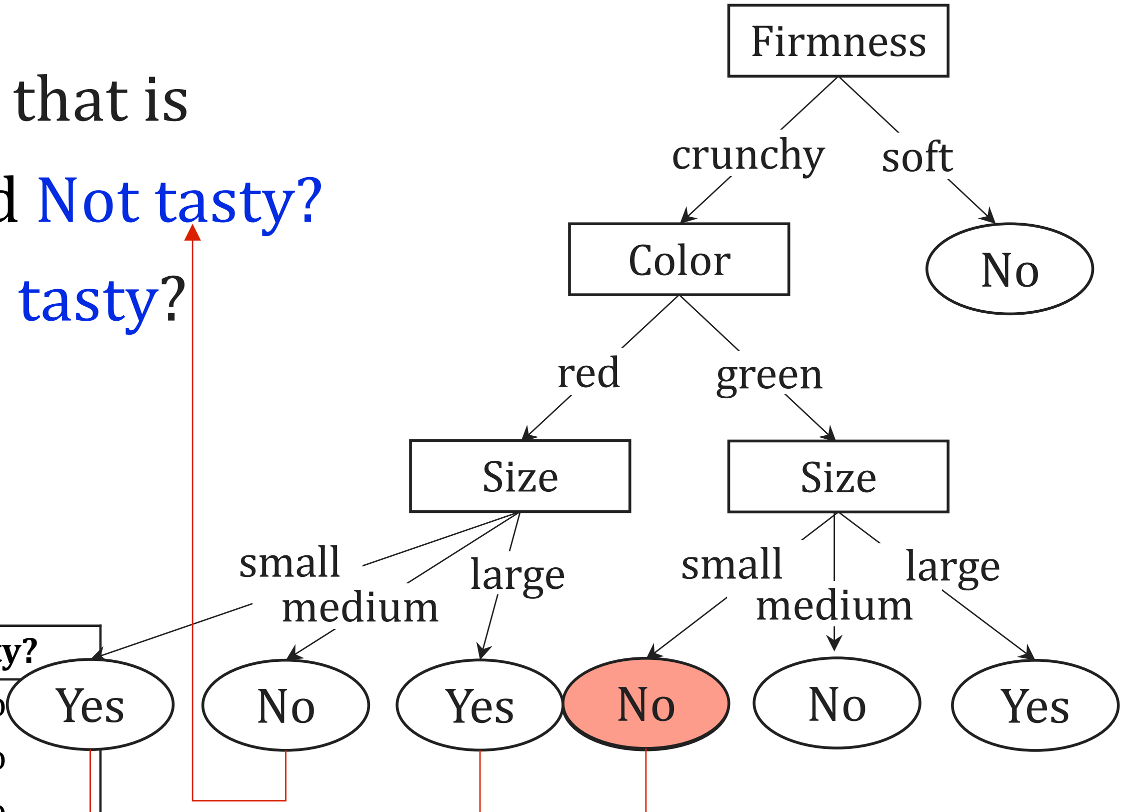
$$h(x) = \begin{cases} y & \text{if } (x,y) \in S \\ \text{flip a coin} & \text{if haven't seen } x \end{cases}$$

Question: When does overfitting happen?

Overfitting in Decision Trees

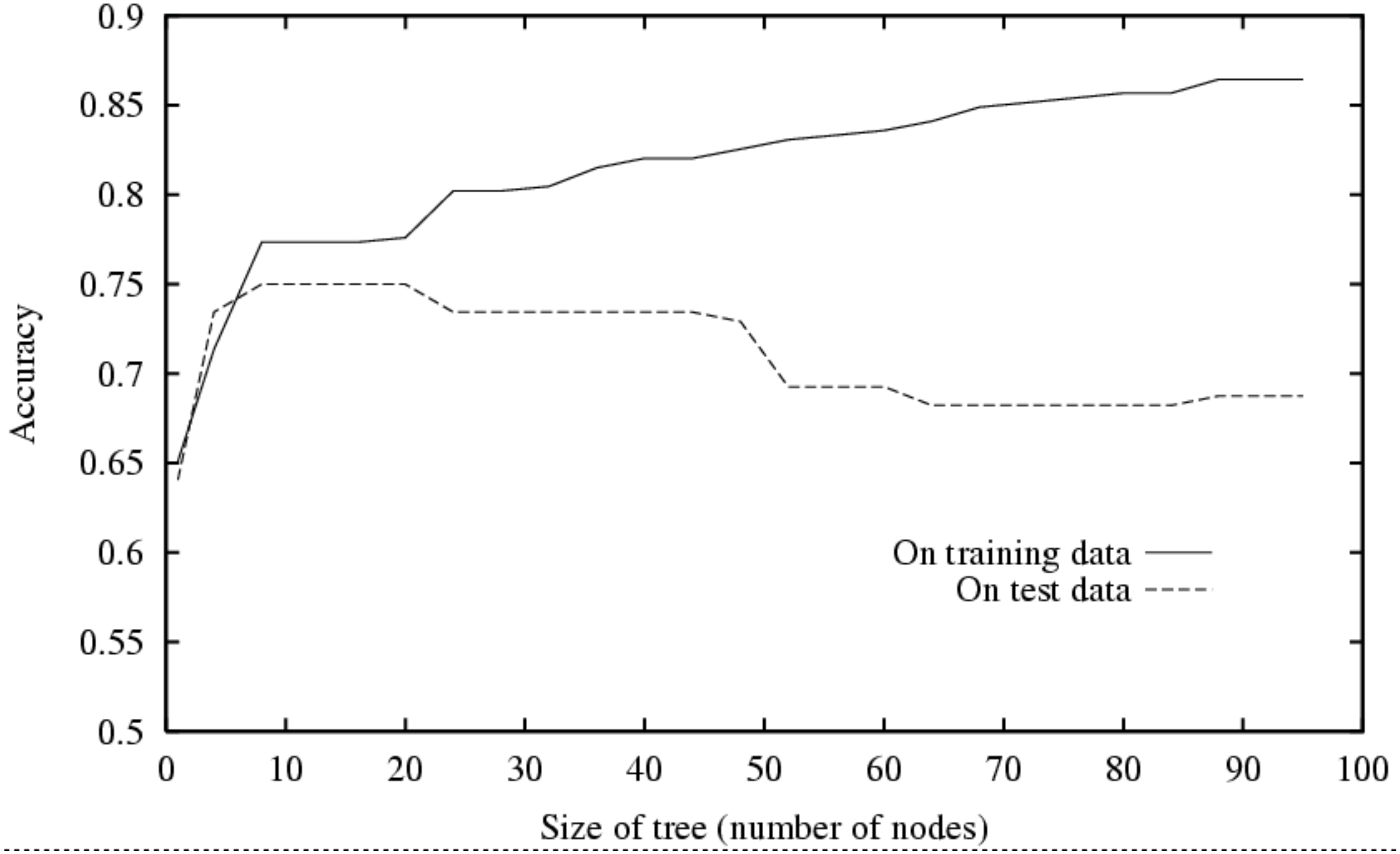
What happens if we an apple that is

- (red, medium, crunchy) and Not tasty?
- (green, small, crunchy) and tasty?



Apple	Color	Size	Firmness	Tasty?
#1	Red	Medium	Soft	No
#2	Green	Small	Crunchy	No
#3	Red	Medium	Soft	No
#4	Red	Large	Crunchy	Yes
#5	Green	Large	Crunchy	Yes
#6	Green	Small	Crunchy	No
#7	Red	Small	Soft	No
#8	Red	Small	Crunchy	Yes
#9	Green	Medium	Soft	No
#10	Red	Small	Crunchy	Yes

Overfitting in Decision Trees



Need for Inductive Bias

Recall: h_S that memorizes S fully (and flips a coin for any x that doesn't appear in the samples overfits.)

Avoid overfitting:

- Should we use a hypothesis space that includes all possible functions., i.e., $H = 2^X$?
  → Restrict hypothesis space, e.g., ANDs, ORs, Decision Lists, …

- Other assumptions?

Inductive Bias in ID3

ID-3: The top-down Induction on DTs using entropy. Make a leaf node

→ if all samples have the same label.

→ if there is no unused feature. Go with the majority label.

Recall: Decision trees are very expressive.

→ How large is the set of DTs on instance space X.

→ Is there no bias?

Inductive bias in ID3 is a *preference* for some hypotheses (fewer nodes), not a restriction to a hypothesis space.

ML Tools for dealing with overfitting

Statistical Learning Theory:

- For which hypothesis sets is learning (without overfitting) possible?
- How large a training set do I need to avoid overfitting?
- **We will learn this later in the course!**

Occam's Razor:

- The law of briefness!
- All things equal, simpler explanations are better.

Example: Two trees fell down during a windy night.

- The wind knocked them down?
- Two meteorites each took one tree down and, after striking the trees, hit each other removing any trace of themselves?