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 \( f: X \to Y \) that is how instance are labeled in life.

Training Data:
A set \( S \) of labeled pairs \( (x_i, f(x_i)) \in X \times Y \) that we have seen before.

Hypothesis space:
A predetermined set \( H \) of functions in which we look for \( h: X \to Y \).

Given a large enough number of training examples, and given a hypothesis space \( H \), learn a hypothesis \( h \in H \) that approximates \( f \).

How does performance of \( h \) on \( X \) translates to unseen instances.

Inductive Learning

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\(^\ast\) Homework \( P(X, Y) \) indicates the probability that a homework with features \( X \) will receive an A\(^\ast\) label \( Y \).

\[
P(X = \text{(Complete, Yes, Yes, Clear, No)}, \text{Yes})
\]

- Tasty Apple \( P(X, Y) \) indicates the probability that an apple with features \( X \) has a taste label \( Y \).

\[
P(X = (\text{A, Red, Medium, Crunchy, Yes)})
\]

Learning as Prediction

Chain rule: Sampling as a two step procedure

\[
P(X, Y) = P(X) \times P(Y | X)
\]

\( P(X) \): Prob. the world produces instance with representation \( X \)

\( P(Y | X) \): Prob. of seeing label \( Y \) on instance \( X \).

**Example 1:** Prob. the teacher assigns A\(^\ast\) to homework \( X \). Deterministic label.

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

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

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

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 \( X_i \) 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.

- \( P(X_i = x_i, Y_i = y_i) = P(X_i = x_i) P(Y_i = y_i) \) for all \( i \) and \( j \).

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

- What about for apples?

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

\[
\Pr(S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}) = \prod_{i=1}^{m} P(X = x_i, Y = y_i)
\]
Sample & Generalization Errors

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

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

Sample error of hypothesis $h$ on samples $S = \{(x_i, y_i)\}$,..., $\{(x_m, y_m)\}$, denoted by $\err_S(h)$ is

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

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

$$\err_P(h) = E_{(x, y) \sim P}(\Delta(h(x), y))$$

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 150,000 common shares of NECO Enterprises Inc. at 26 cents each. He said the unsuccessful completion of the offer, plus shares he already owns, would give him 50.5
doing NECO Enterprises.

SALANT CORP 1ST QTR FEB 28 NET

Oper net profit 216,000 vs loss 401,000. Sales 21.4

“...and many more

Overfitting

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

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

Question: Does $h$ overfit on samples $S = \{(x_i, f(x_i))\}$,..., $\{(x_m, f(x_m))\}$?

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

Question: When does overfitting happen?

Decision Tree for “Corporate Acq.”

Learned tree:
  • has 437 nodes
  • is consistent
  • $\err_{\text{train}}(h) = 0$

Accuracy of learned tree:
  • $\err_{\text{test}}(h) = 0.11$

Note: word stems expanded for improved readability.

Overfitting in Decision Trees

What happens if we an apple that is
  • (red, medium, crunchy) and Not tasty?
  • (green, small, crunchy) and tasty?

Overfitting in Decision Trees

Overfitting in Decision Trees

Overfitting in Decision Trees

Overfitting in Decision Trees
Overfitting in Decision Trees

![Graph showing accuracy vs. size of the sample]

Need for Inductive Bias

**Recall:** $h$ 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 brie ness!
- 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?