CSE 417T: Introduction to Machine Learning

Lecture 2: Generalization

Henry Chai
08/29/19
Course Information:

- Website: classes.cec.wustl.edu/~SEAS-SVC-CSE417T/
- Piazza for questions (signup with Wash U email)
- Gradescope for submitting homework (signup with 9J6YJ3)
- Canvas for grades (should already be signed up)
- Textbooks:
  - Learning From Data (AML), http://amlbook.com/
• First homework will be posted tomorrow (8/30)!
  • No more than 2 late days
  • Feel free to discuss homework with other students but you must write your own solutions and cite all external sources (including other students)
  • Bonus question:
    • Write a multiple choice question related to the content of the homework
    • 1 point for each of: completion, correctness (technically and grammatically), relevance, appropriateness (difficulty) and creativity
Writing good multiple choice questions

- Good multiple choice questions can be answered without looking at the options
- Answers are all about the same length and grammatical structure
- Wrong answers must be clearly and justifiably inferior
  - Avoid absolutes (never/always) and fuzzy language (rarely/usually)
  - All/none of the above are generally pretty bad
Types of Learning

• Supervised Learning
  • Training data is (input, output)
  • Examples: linear/logistic regression, support vector machines, neural networks
  • Variants: active learning and online learning

• Unsupervised Learning
  • Training data is (input)
Types of Learning
Types of Learning

- **Supervised Learning**
  - Training data is (input, output)
  - Examples: linear/logistic regression, support vector machines, neural networks
  - Variants: active learning and online learning

- **Unsupervised Learning**
  - Training data is (input)
  - Examples: clustering, principal component analysis, outlier detection

- **Reinforcement Learning**
  - Training data is (input, action, score)
  - Examples: Q-learning, temporal difference learning
Types of Learning

Source: https://www.xkcd.com/242/
• Supervised Learning (this class!)
  • Training data is (input, output)
  • Examples: linear/logistic regression, support vector machines, neural networks
  • Variants: active learning (CSE 515T) and online learning

• Unsupervised Learning (CSE 517A)
  • Training data is (input)
  • Examples: clustering, principal component analysis, outlier detection

• Reinforcement Learning (CSE 511A)
  • Training data is (input, action, score)
  • Examples: Q-learning, temporal difference learning
Puzzle

\[ f(x) = +1 \]

\[ f(x) = -1 \]

\[ f(x) = ??? \]
What do you think the label of this example is?

\[ f(x) = +1 \]
\[ f(x) = -1 \]
\[ f(x) = ??? \]
An Answer

\[ h(x) = \begin{cases} 
+1 & \text{if top left is white} \\
-1 & \text{otherwise} 
\end{cases} \]

\[ h \left( \begin{array}{ccc}
\text{white} & \text{white} & \text{white} \\
\text{black} & \text{black} & \text{black} \\
\text{black} & \text{black} & \text{black} \\
\end{array} \right) = -1 \]
An Answer

\[ h(x) = \begin{cases} 
+1 & \text{if } x \in \{ \ \} \\
-1 & \text{otherwise} 
\end{cases} \\
\downarrow \\
h(\begin{array}{c}
\end{array}) = -1
\]
An Answer

\[ h(x) = \begin{cases} 
+1 & \text{if } x \in \{ \begin{array}{c} \text{patterns} \end{array} \} \\
-1 & \text{otherwise} 
\end{cases} \]

\[ h \left( \begin{array}{c} \text{pattern} \end{array} \right) = +1 \]
Recall

**Unknown target function**

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

Training data

\[ \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

Learning Algorithm

\[ \mathcal{A} \]

Hypothesis Set

\[ \mathcal{H} \]

Learned Hypothesis

\[ g : \mathcal{X} \rightarrow \mathcal{Y} \]
Break: Review

- Is learning feasible?
  - If you demand absolute certainty in your predictions, then no...

- Does this mean we’re doomed?!?!
Analogy

\( \mu \) = fraction of red marbles in bin

\( \nu \) = fraction of red marbles in sample

Does \( \nu \) say anything about \( \mu \)?
Hoeffding’s Inequality

- $\mu = \text{fraction of red marbles in bin}$
- $\nu = \text{fraction of red marbles in a sample of size } n$

$$P\{|\nu - \mu| > \epsilon\} \leq 2e^{-2\epsilon^2n}$$

- As $n$ increases, RHS decreases
- As $\epsilon$ decreases, RHS increases
= input space ($\mathcal{X}$)

= point in the input space ($\vec{x}$)

= training data ($\mathcal{D}$)

= a point classified correctly by a specified hypothesis $h$

= a point classified incorrectly by a specified hypothesis $h$

$\nu = \text{fraction of training data classified incorrectly by } h$

$\mu = \text{fraction of points in the input space classified incorrectly by } h$
= input space \((\mathcal{X})\)

= point in the input space \((\mathcal{X})\)

= training data \((\mathcal{D})\)

= a point classified correctly by a specified hypothesis \(h\)

= a point classified incorrectly by a specified hypothesis \(h\)

\(\nu = \) fraction of training data classified incorrectly by \(h\) \((E_{in}(h))\)

\(\mu = \) fraction of points in the input space classified incorrectly by \(h\) \((E_{out}(h))\)

\[ P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n} \]
Formal Setup

Unknown target function

\[ f: \mathcal{X} \rightarrow \mathcal{Y} \]

Hypothesis Set

\( \mathcal{H} \)

Learning Algorithm

\( \mathcal{A} \)

Training data

\( D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \)

Probability Distribution

\( \mathcal{P} \) on \( \mathcal{X} \)

Learned Hypothesis

\( g: \mathcal{X} \rightarrow \mathcal{Y} \)
Connection to Learning Validation

= input space ($\mathcal{X}$)

= point in the input space ($\mathbf{x}$)

= training data ($\mathcal{D}$)

= a point classified correctly by a specified hypothesis $h$

= a point classified incorrectly by a specified hypothesis $h$

$\nu$ = fraction of training data classified incorrectly by $h$ ($E_{in}(h)$)

$\mu$ = fraction of all possible data classified incorrectly by $h$ ($E_{out}(h)$)

$$P\{|E_{in}(h) - E_{out}(h)| > \epsilon\} \leq 2e^{-2\epsilon^2 n}$$
Formal Setup

Unknown target function

\[ f: \mathcal{X} \rightarrow \mathcal{Y} \]

Probability Distribution

\[ \mathcal{P} \text{ on } \mathcal{X} \]

Training data

\[ \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

Learning Algorithm

\[ \mathcal{A} \]

Hypothesis Set

\[ \mathcal{H} \]

Learned Hypothesis

\[ g: \mathcal{X} \rightarrow \mathcal{Y} \]
Formal Setup

Unknown target function

\[ f: \mathcal{X} \rightarrow \mathcal{Y} \]

Probability Distribution

\[ \mathcal{P} \text{ on } \mathcal{X} \]

Training data

\[ \mathcal{D} = \{ (x_1, y_1), \ldots, (x_n, y_n) \} \]

Learning Algorithm

\[ \mathcal{A} \]

Hypothesis Set

\[ \mathcal{H} \]

Learned Hypothesis

\[ g = \underset{h \in \mathcal{H}}{\text{argmin}} L(h) \]
Another Analogy

• If you toss a fair coin 20 times, the probability that it comes up heads 20 times is $2^{-20} \approx 1e-6$

• If you toss $2^{20}$ fair coins 20 times each, the probability that at least one coin comes up heads 20 times is...

... $\approx 1 - \frac{1}{e} \approx 0.63$
Another Analogy

- Take any coin that came up all heads and apply Hoeffding’s inequality

\[ P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2e^{-2\epsilon^2 n} \]

\[ P\{|\text{Fraction of tails in 20 trials} - \text{Fraction of tails in } \infty \text{ trials}| > \epsilon\} \leq 2e^{-40\epsilon^2} \]

\[ P\{|0 - P\{\text{this coin coming up tails}\}| > \epsilon\} \leq 2e^{-40\epsilon^2} \]

\[ P\left\{P\{\text{this coin coming up tails} > \frac{1}{4}\}\right\} \leq 2e^{-2.5} \approx 0.15 \]
Hoeffding’s Inequality (Corrected)

• Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, \ldots, h_m\}$ and $g \in \mathcal{H}$

\[
P\{|E_{in}(g) - E_{out}(g)| > \epsilon\}
\]

\[
\leq P \left\{ \bigcup_{j=1}^{m} |E_{in}(h_j) - E_{out}(h_j)| > \epsilon \right\}
\]

\[
\leq \sum_{j=1}^{m} P\{|E_{in}(h_j) - E_{out}(h_j)| > \epsilon\}
\]

\[
\leq \sum_{j=1}^{m} 2e^{-2\epsilon^2 n} = 2(m)e^{-2\epsilon^2 n}
\]
Hoeffding’s Inequality (Corrected)

- Suppose $\mathcal{H}$ is finite i.e. $\mathcal{H} = \{h_1, ..., h_m\}$
- $E_{in}(g) = \text{in-sample error of best hypothesis in } \mathcal{H}$
- $E_{out}(g) = \text{out-of-sample error of best hypothesis in } \mathcal{H}$

- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2(m)e^{-2\epsilon^2n}$

- As $n$ increases, RHS decreases
- As $\epsilon$ decreases, RHS increases
- As $m$ increases, RHS increases