Tuesday, 9/15/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 7

Announcements

- Homework 2 (Week 3) is due Friday 9/18.
- Starting this week, lecture topics are in the AML book.

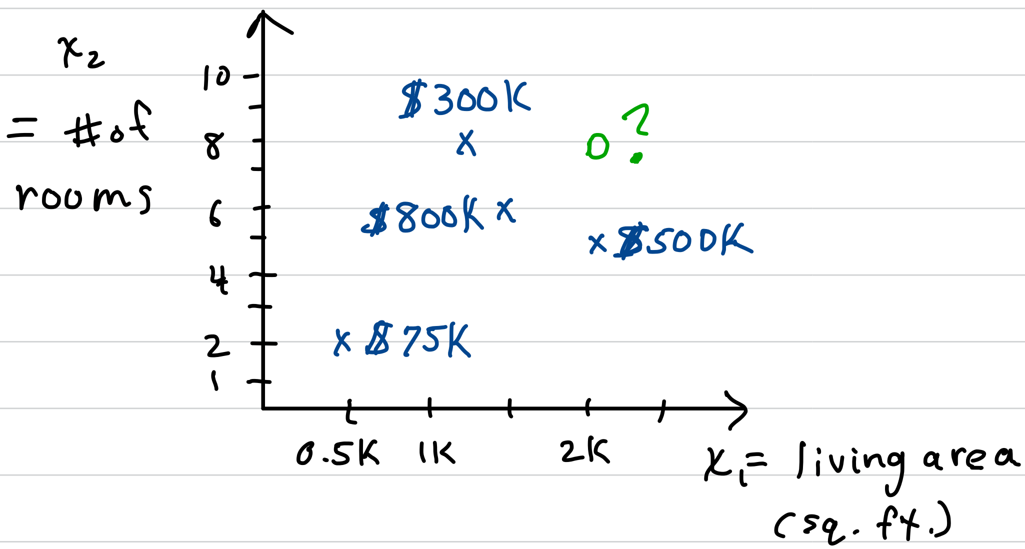
Today's Lecture

- Feasibility of learning (part 1)
  - Understanding ML and its feasibility
  - Generalization error
  - Marbles and bins
  - Hoeffding inequality
    - Single hypothesis
    - *M* hypotheses

Feasibility of Learning

How or why can we expect a machine to generalize from training data $\mathcal{D}$?

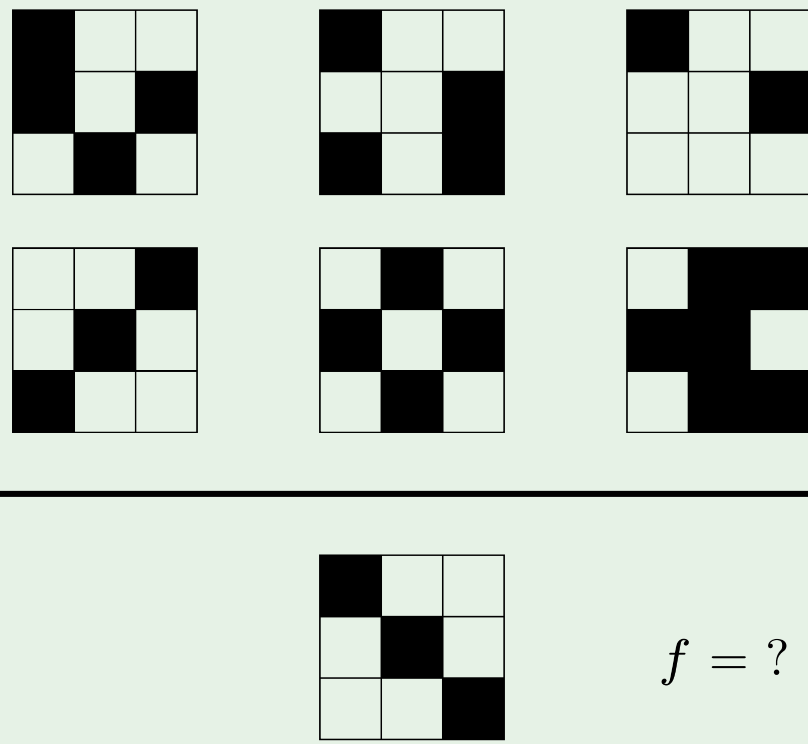
Housing price example:



x trainig data

o unknown

A Learning puzzle



[Fig. 1.7 of Abu-Mostafa, et al., Learning from Data (AML)]

What can make machine learning feasible?

Apparently the following can help:

- Assumptions
- Prior knowledge
- Appropriate hypothesis set (model)
- Enough data
- Appropriate features

Suggestions:

- Models
- Larger N

Generalization Error

Let $E_{in}(h) = E_{\mathcal{D}}(h)$ = in-sample error (error on dataset $\mathcal{D}$)

For classification problems, let:

$$E_{\mathcal{D}}(h) = \frac{1}{N}\sum_{n=1}^{N} [\![h(\underline{x}_n) \neq f(\underline{x}_n)]\!]$$

hypothesis prediction $h(\underline{x}_n) = \hat{f}(\underline{x}_n)$ or $\hat{y}(\underline{x}_n)$

true target function $f(\underline{x}_n) = y_n$

$\therefore E_{\mathcal{D}}(h)$ = % of points in $\mathcal{D}$ that are misclassified by h.

$\mathcal{D}$ can be $\mathcal{D}_{Tr}$, $\mathcal{D}_{Val}$, or $\mathcal{D}_{Te}$; then $N = N_{Tr}, N_{Val}, N_{Te}$.

Let $E_{out}(h)$ = out-of-sample error (probability of error over all $\underline{x}$)

$$E_{out}(h) = P[h(\underline{x}) \neq f(\underline{x})] \quad \text{(for classification problems)}$$

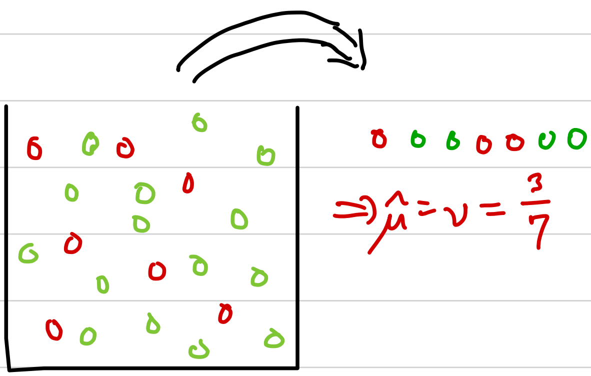
$\boxed{\text{We can measure } E_{\mathcal{D}}\text{; we want to know } E_{out}}$ $\leftarrow$ Fundamental issue in ML

Marbles and bins, μ and ν

We want to compare $E_{\mathcal{D}}(h)$ with $E_{out}(h)$, when $E_{out}(h)$ is unknown.

Measure $E_{\mathcal{D}}(h)$ from $\mathcal{D}$:

↳ a sample of data points drawn from input space $\mathcal{X}$ according to $p(\underline{x})$.



$\Rightarrow \hat{\mu} = \nu = \frac{3}{7}$

Draw a marble

Let $\mu = P(\text{red})$

then $1 - \mu = P(\text{green})$

μ is deterministic but unknown to us.

How can we find or estimate μ?

Draw N marbles independently (with replacement) z_i, $i = 1, 2, \cdots, N$.

Estimate $\hat{\mu} = \dfrac{\#\text{red marbles}}{N} \triangleq \nu$

How accurate is our estimate?

Hoeffding Inequality

$$P\left[\,|\nu-\mu|>\epsilon\,\right]\le 2e^{-2\epsilon^2 N} \qquad \text{for any } \epsilon>0$$

our tolerance

size of $\mathcal{D}$
(sample size, # of marbles drawn

Note dependence of R.H.S. on ϵ, N.

Relate this to ML $(\underline{w}=\underline{w}^{(t)})$

2-class classification problem

Hypothesis: $h(\underline{x}) = [\![\, \tilde{\underline{w}}^T \underline{x} \ge 0 \,]\!]$, $\tilde{\underline{w}}$ is given

Target fcn. is $f(\underline{x})$ (true class label)

ML	Marbles & bins
1. Pick a data point from $\mathcal{X}$ according to $p(\underline{x}) \rightarrow \underline{x}_i$.	1. Draw a marble from the bin
If $h(\underline{x}_i) = f(\underline{x}_i)$, then $\underline{x}_i$ is classified correctly by h	Marble is colored green (probability $1-\mu$)
If $h(\underline{x}_i) \neq f(\underline{x}_i)$, then $\underline{x}_i$ is misclassied by h	Marble is colored red (probability μ)
2. Pick N data points, i.i.d. from $\mathcal{X}$ $\Rightarrow$ dataset $\mathcal{D}$ (sample)	2. Draw N marbles from bin, with replacement
We can calculate $E_{\mathcal{D}}(h)$	$= \nu$ = % of drawn marbles that are red
We want to know or bound $E_{out}(h)$	$= \mu = P[\text{marble is red}]$

$\Longrightarrow$ Use Hoeffding Inequality

Feasibility of Learning and Hoeffding Inequality

(1) $$\boxed{P\left[\,\left|E_{\mathcal{D}}(h) - E_{out}(h)\right| > \epsilon\,\right] \leq 2e^{-2\epsilon^2 N} \quad \text{for any } \epsilon > 0}$$

Note: $\mathcal{D}$ must be drawn at random, i.i.d., according to $p(\underline{x})$.

Procedure for Hoeffding Inequality to be valid:

1. Specify h (determines color of marbles)
2. Draw $\mathcal{D}$ (draw N marbles from bin)
3. Calculate $E_{\mathcal{D}}(h)$; get bound on $E_{out}(h)$ using (1).

ML paradigm:

1. Collect dataset $\mathcal{D}$ (and clean the data)
2. Construct hypothesis set $\mathcal{H}$.
3. Train to find h_g (best hypothesis from $\mathcal{H}$)
4. Calculate $E_{Tr}(h_g)$ or $E_{Test}(h_g)$
5. Want to know $E_{out}(h_g)$.

→ Order is different.

→ Resolve by applying Hoeffding Ineq. <u>before</u> choosing h_m. (one hypothesis)

AML shows that:

$$P\left[\left| E_{\mathcal{D}}(h_g) - E_{out}(h_g) \right| > \epsilon \right] \le P\left[\bigcup_{m=1}^{M} \left(|E_{\mathcal{D}}(h_m) - E_{out}(h_m)| > \epsilon \right) \right]$$

$$\le \sum_{m=1}^{M} \underbrace{P\left[|E_{\mathcal{D}}(h_m) - E_{out}(h_m)| > \epsilon \right]}_{\le 2e^{-2\epsilon^2 N}}$$

So:

(2)
$$\boxed{P\left[\left| E_{\mathcal{D}}(h_g) - E_{out}(h_g) \right| > \epsilon \right] \le 2Me^{-2\epsilon^2 N}}$$

$M = |\mathcal{H}|$ = # of hypotheses in $\mathcal{H}$

A loose bound. (but sometimes useful).

→ We will get a tighter bound later.