Key for Moodle: MSFML25

Tutorials: Thursdays, 2:15, Ther. 39, B133

Exam: oral (likely), ~ 10-14 days after July 26,
2-4 weeks.

Lecture notes uploaded.

ML: computational method to convert experience ⤳ expertise.

- Experience: past data.
- Expertise: prediction of future outcomes

Std learning tasks: classificatn, regression, clustering, dimensionality reduction/ manifold learning - find lower dimnl. rep of data while preserving their props.

This course:

1. Thtcl foundatns: & statistical learning thy PAC Learning Framework, Rademacher complexity, VC-dimension
2. Analysis of ML methods & algos (w/ apps. of pt 1).
   - SVM & Kernel methods
   - Boosting
   - Logistic regrssn
   - SGD
   - NNs (DL)
   - May be part 2.

Basic ML Setup.

Input space X, eg $\mathbb{R}^n$, $[0,1]^2$, ...

Output space Y, eg $\{0,1\}$, $\mathbb{R}$, ...

Training data (x_i, y_i), $i = 1, \dots, m$ (labeled data)
[or x_i, $i = 1, \dots, m$ (unlabeled)

Hypothesis class $\mathcal{H}$: a set of functions $h: X \to Y$.

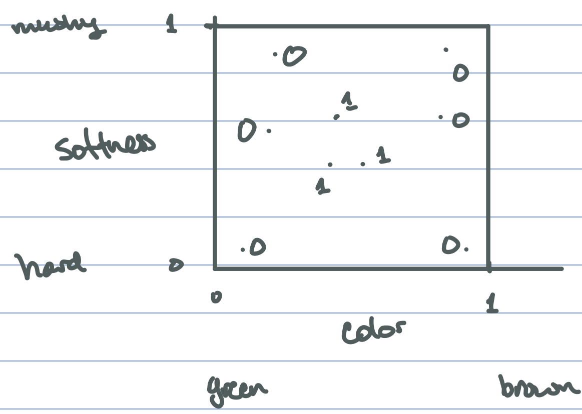
What class of functions to learn? Why do we need to choose it?

Suppose you want to predict whether a papaya is tasty or not.

Based on experience w/ other fruits, we assume that taste depends on color & softness.

Assume we measure both on an interval $[0,1]$.

$$\rightsquigarrow \mathcal{X} = [0,1]^2$$



0 ≙ not tasty
1 ≙ tasty

$\rightsquigarrow$ Try a couple fruits, measure softness & color & label w/ 0 or 1.

Goal: find a function $h : [0,1] \to \{0,1\}$ s.t.

$$h(a,b) = \begin{cases} 1 & \text{if papaya w/ color } a \text{, softness } b \text{ is typically tasty,} \\ 0 & \text{if ——— " ——— not tasty.} \end{cases}$$

The pairs $(x_i, y_i) = ((a_i, b_i), y_i)$ are distributed according to some unknown distribution D, representing environ.

Given $(x_1, y_1), \ldots, (x_n, y_n) \in [0,1]^2 \times \{0,1\}$, how should we choose h?

h should satisfy $h(x_i) = y_i$ for (at least almost) all (x_i, y_i).

There may be many such functions.

Eg. $$h(x) = \begin{cases} 1 & \text{if there is a training eg } (x_i, y_i) \text{ w/ } y_i = 1 \\ 0 & \text{o/w} \end{cases}$$

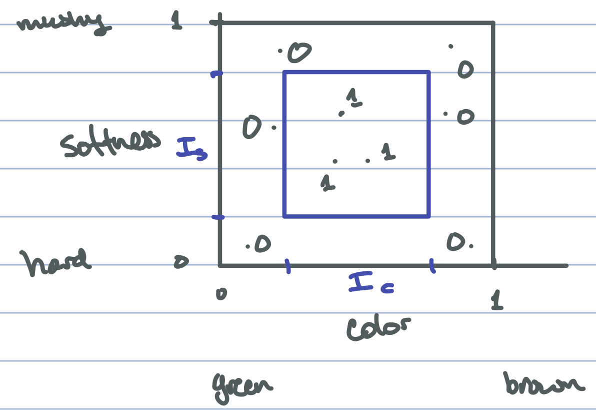
Does not seem realistic. ($\to$ overfitting).

Then $h(x_i) = y_i \ \forall i$ (x_i distinct).

Other possibility (more realistic): assume there are intervals $I_c \subset [0,1]$ and $I_s \subset [0,1]$ s.t. papayas w/ color value $\in I_c$, softness value $\in I_s$ are (typically) tasty, other papayas not tasty.

$$h(a,b) = \begin{cases} 1 & \text{if } (a,b) \in I_c \times I_s, \\ 0 & \text{o/w}. \end{cases} \rightsquigarrow h = \mathbb{1}_{I_c \times I_s}.$$

Want to define class H from which we choose h, eg $\mathcal{H}$ could be the set of characteristic functions of axis-aligned rectangles.



How to choose such a rectangle?

Choosing a suitable class $\mathcal{H}$ requires preknowledge about the problem.

In general, $\mathcal{H}$ should not be the class of all functions

$$[0,1]^2 \to \{0,1\}$$

1. Supervised.

Learner receives labeled data (x_i, y_i) $\rightsquigarrow$ tries to find function
$i = 1, \dots, m$

$$h(x_i) \approx y_i,$$

$i = m+1, m+2, \dots$

Example: $x_i = ([0,255] \cap \mathbb{Z})^{n_1 \times n_2}$, $y_i = \{1, 0\}$ cat in img or not

$3 n_1 \times n_2 \rightsquigarrow$ if you include color channels.

2. Unsupervised.

Learner receives unlabeled data x_i, $i=1,\dots,m$ $\rightsquigarrow$ understd structure of data points. They may be contained in a much smaller submanifold / subspace, or cluster.

3. Semisupervised

Learner $\leftarrow (x_i, y_i)$; $i=1,\dots,m$; unlabeled x_i, $i=m+1,\dots,N$. Predict labels of unlabeled data; helps learn struct of data.

4. Online learning.

Mult. training rounds. Receive x_i, predict $\hat{y}_i$ $\rightsquigarrow$ receive real label y_i $\rightsquigarrow$ incur loss $\ell(y_i, \hat{y}_i)$. Goal: minimize cumulative loss

$$\sum_{i=1}^{m} \ell(y_i, \hat{y}_i).$$

5. RL

Mixed training/testing phases. Learner $\rightarrow$ actions $\rightsquigarrow$ receives immediate reward/loss. Goal: maximize reward over course of actions.

$\hookrightarrow$ explore vs. exploit dilemma.

explore: get info abt environ, unexplored action

exploit: known action, exploit already collected information

6. Active learning

Learner can decide on data points x_i to query the label y_i.

Eg, scientific experiments, oil exploration.

Hope: better predictions / less samples.

PAC Learning Framework.

Consider supervised learning scenario for binary classifn

X: input, $Y = \{0,1\}$ output, training data

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$$

Assumptions.

1. x_i, $i = 1, \ldots, m$ drawn iid, according to some unknown prob. dist. D on X.

2. Labels are given $y_i = f(x_i)$, $i = 1, \ldots, m$ for some map $f : X \to Y$.

Goal: find f (at least approximately).

More generally (later): assume (x_i, y_i) drawn iid from D on $X \times Y$.

Learner's output: a prediction rule

$$h : X \to Y,$$

ideally $h = f$.

Learner selects hypothesis from

$$\mathcal{H} \subset \{g : X \to Y\}$$

and f may or may not be contained in $\mathcal{H}$.

Assumptions on measurable spaces: for msbl spaces $X, Y, Z, \ldots$

1. If space is finite/countably infinite, it is equipped w/ the σ-algebra consisting of all subsets of the space.

2. Otherwise, assume it is a metric space, complete & separable, equipped w/ corresponding Borel σ-alg.

3. For products $X \times Y$, carries the product σ-alg of X and Y.

Generalization Error. For $h : X \to Y = \{0,1\}$, the generalization error or risk of h is:

$$R(h) = \underset{x \sim D}{\mathbb{P}}(h(x) \neq f(x)) = \mathbb{E}[\mathbb{1}_{\{x | h(x) \neq f(x)\}}]$$

$\rightsquigarrow$ note $\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$

"prob. of a set"

$= \int_\Omega \mathbb{1}_A \, dP(\omega)$

$= \int_A 1 \, dP(\omega)$

- Above, we assumed a fixed labeling function $f: X \to Y = \{0,1\}$ and probability distribution $\mathcal{D}$ on X.