Mathematical Foundations of ML (CS 4785/5783)
Lecture 1

Setting up the Learning Problem

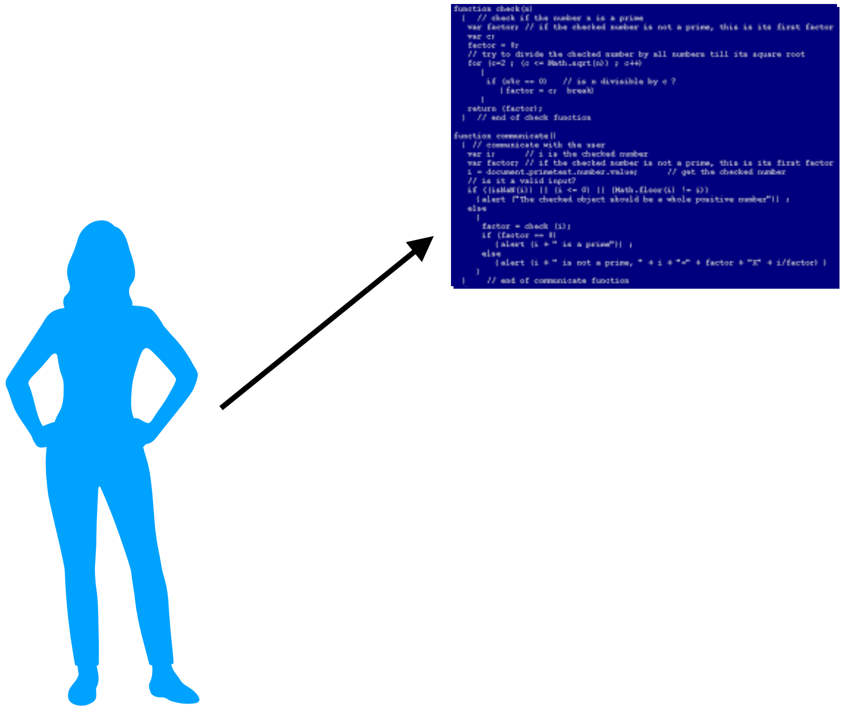
http://www.cs.cornell.edu/Courses/cs4783/2023fa/notes01.pdf

Traditional Computer Science

Task
Eg. Sorting

Input
Eg: 2, 4, 3, 8, 7

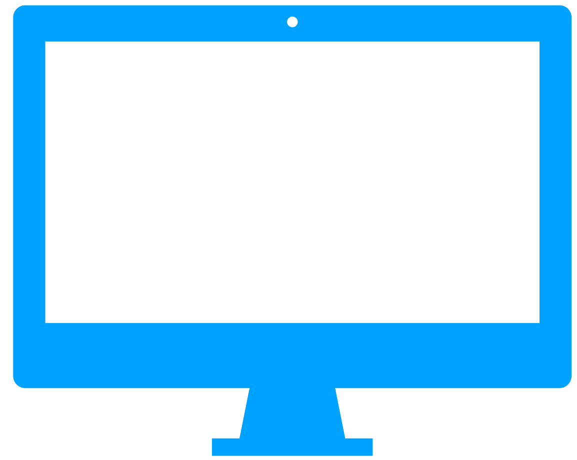
Output
Eg: 2, 3, 4, 7, 8



Machine Learning

Task
Eg. Object recognition

Input
New input instance



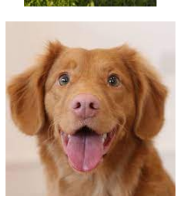
Output
Prediction of label

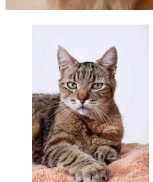
“cat”

Input
A set of input/output pairs

 **, “cat”**

 **, “dog”**

 **, “dog”**

 **, “cat”**

What is Machine Learning

"The field of study that gives computers the ability to learn without being explicitly programmed" - Arthur Lee Samuel

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" - Tom Mitchell

Learning Problem : Basic Notation

- Input space/ feature space : $\mathcal{X}$
  (Eg. bag-of-words, n-grams, vector of grey-scale values, user-movie pair to rate)

- Output space/ label space $\mathcal{Y}$
  (Eg. $\{\pm 1\}$, $[K]$, $\mathbb{R}$-valued output, structured output)

- Loss function : $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$
  (Eg. $0-1$ loss $\ell(y', y) = \mathbf{1}\{y' \neq y\}$, sq-loss $\ell(y', y) = (y - y')^2$), absolute loss $\ell(y', y) = |y - y'|$

  Measures performance/cost per instance (inaccuracy of prediction/ cost of decision).

Two Scenarios

Universe of instances

U

$f_{i^*}\Bigg($  $\Bigg)$ **= “not cat”**

$f_{i^*}\Bigg($  $\Bigg)$ **= “cat”**

i* in [N] is unknown

Amongst set of models $\{f_1, \ldots, f_N\}$

There is the perfect model f_{i^*}

Two Scenarios

Scenario I

Universe of instances

U

Draw n instances from the universe at random and label them

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

S is called training set!

x_i's are images taken from the universe

$y_i = f_{i^*}(x_i)$

Learning algorithm has access to the models $\{f_1, \ldots, f_N\}$

Goal: return a model with small classification error

Scenario I

What should the learning algorithm be?

What kind of guarantee can we provide on its error?

How does our guarantee (bound) on error depend on N the number of models, on n the number of samples we drew?

Scenario I

Algorithm: return any classifier that is consistent with S

$$\text{Return: } \hat{f}_S \in \{f_i : \forall t \in [n], f_i(x_t) = y_t\}$$

Error bound:

For any $\delta > 0$, with probability at least $1 - \delta$ over draws of S,

$$P(\hat{f}_S(x) \neq y) \leq \frac{\log(N/\delta)}{n}$$

PAC: Probably Approximately Correct

SCENARIO II

Universe of instances

U

On each round t:

Email x_t is composed, possibly by spammer!

System classifies email as $\hat{y}_t$

True label $y_t = f_{i^*}(x_t)$ revealed

We get feedback every round. But spammer can pick next email.

Goal: Make as few mistakes as possible.

Scenario II

What should the learning algorithm do?

What is the bound on total number of mistakes made?

Scenario II

How about using the same algorithm from scenario 1 for each t (re-run)?

How many mistakes would it make?

Scenario II

Algorithm:

$$\text{Pick } \mathcal{F}_t = \{f_i : i \in [N], \forall s < t, f_i(x_s) = y_s\}$$
$$\text{Set } \hat{y}_t = \text{Majority}(\{f(x_t) : f \in \mathcal{F}_t\})$$

Mistake Bound:

$$\sum_t \mathbf{1}\{\hat{y}_t \neq y_t\} \leq \log_2 N$$

Why?

What is in this Course?

1. Statistical Learning theory
   A. Generalization Error, Training Vs Test loss, Model Complexity
   B. PAC model and VC theory
   C. Rademacher Complexity and Uniform convergence
   D. Role of Regularization in learning, model selection and validation
   E. Algorithmic Stability
2. Online learning
   1. Perceptron
   2. Online experts problem
   3. Online Gradient descent and Mirror descent
3. Boosting
4. Stochastic Optimization and Learning: including understanding Stochastic Gradient Descent
5. Bandit problems: both stochastic and adversarial settings
6. A primer to theory of deep learning: new challenges
7. Computational Learning theory: Computational hardness or learning, proper vs improper learning
8. Societal aspects of ML: Differential Privacy, Right to be Forgotten, Fairness and ML

Grading

- 3% Class participation
- 4 Assignments worth 40% of your grades
- One prelims worth 30% of your grades
- One Term project worth 27% of your grades
- For CS 5783 additional 2 reading assignment + quizzes on them this will be 10% of grade (prelims 25% and Proj 22%). CS 4783 students can also optionally take this.

Rough Timeline

- Assignments: there are tentative and subject to changes
  - HW1: Aug 23, HW2: Sep 13, HW3: Oct 11, HW4: Nov 13
  - Each assignment has roughly a week
- Prelims: Oct 25th in class
- Project: Initial proposal due mid semester (early oct), there will be a project brainstorming lecture in November (Nov 6th tentative). Final report due exam week