10-701: Introduction to Machine Learning

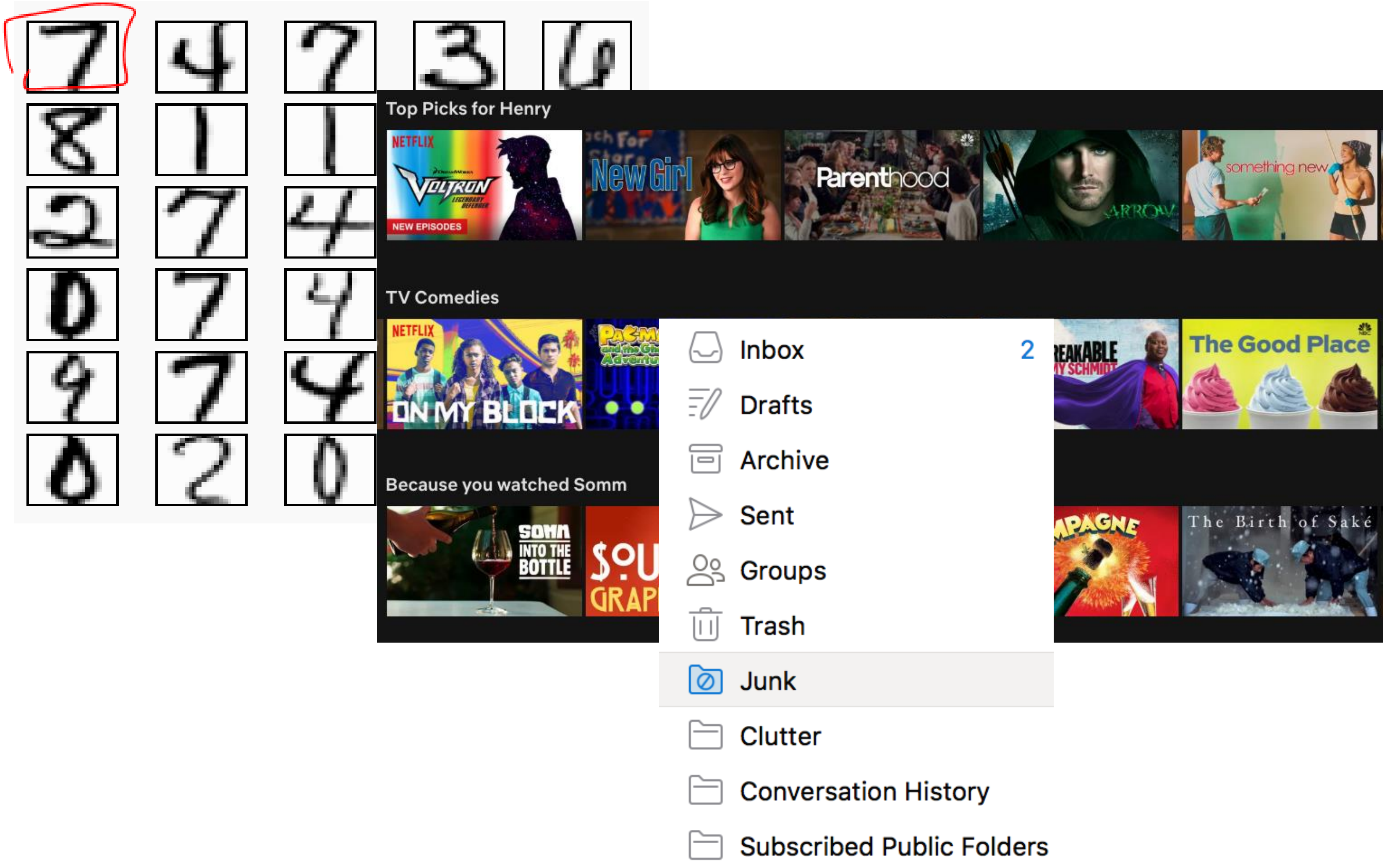
Lecture 1 – Problem Formulation & Notation

Hoda Heidari

8/25/2025

What is Machine Learning?

Machine Learning (A long long time ago…)

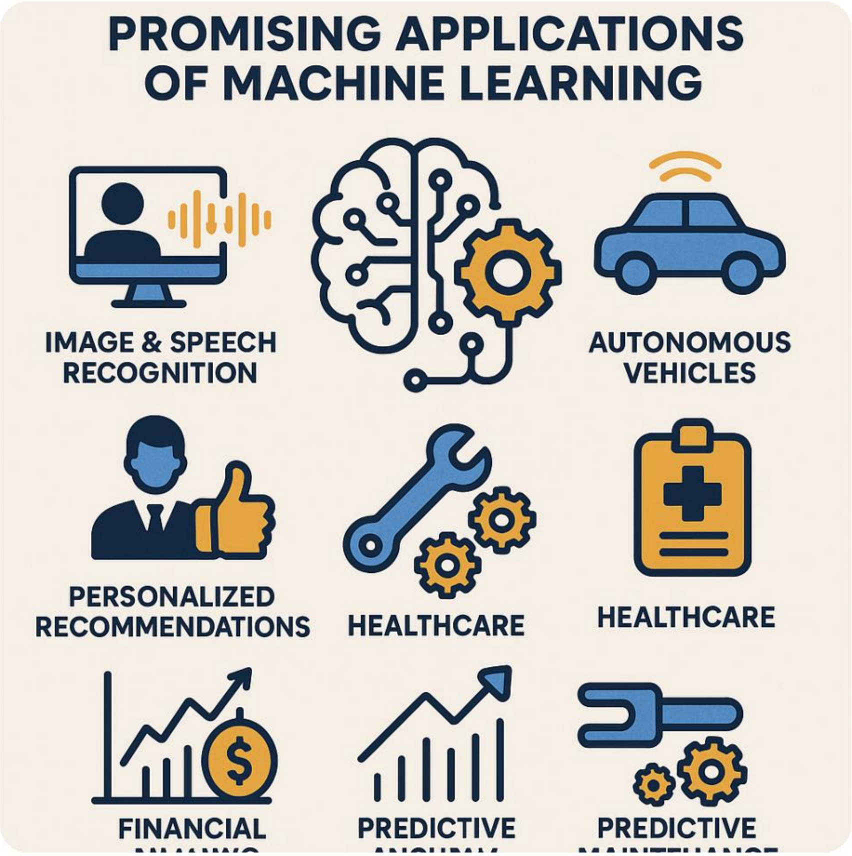


Machine Learning (A short time ago…)

Machine Learning (Now – literally yesterday)

Create an infographic illustrating the most promising applications of machine learning today.

Image created



Source: https://chat.openai.com/

Machine Learning – A Brief Timeline

- Early Foundations (1940s–1960s)
  - **1957:** Frank Rosenblatt develops the *perceptron*, an early neural network for classification.
- Symbolic AI & the First AI Winter (1970s–1980s)
  - Limitations of perceptrons (Minsky & Papert, 1969) and lack of computing power lead to skepticism and reduced funding
- Statistical & Algorithmic Advances (1980s–1990s)
  - **1986:** Rumelhart, Hinton & Williams popularize *backpropagation*, enabling multi-layer neural networks to learn.
  - **1980s–90s:** Emergence of *support vector machines* (SVMs), decision trees, boosting (AdaBoost), and Bayesian methods.
- The Rise of Data & Kernel Methods (1990s–2000s)
  - Explosion of digital data + faster computing power.
  - Kernel methods, ensemble methods, RL
- The Deep Learning Revolution (2010s)
  - **2012:** AlexNet (Krizhevsky, Sutskever, Hinton) wins ImageNet competition using GPUs + deep CNNs, igniting the *deep learning boom*.
  - Reinforcement learning breakthroughs (e.g., DeepMind's AlphaGo in 2016).
- Foundation Models & Generative AI (2020s–present)
  - Rise of *transformers* (Vaswani et al., 2017) revolutionizes NLP (BERT, GPT).
  - Emergence of *foundation models* trained on massive datasets for general-purpose use.
  - Policy, ethics, and responsible AI practices gain prominence due to societal impacts.

What is ~~Machine Learning~~ 10-301/601?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - Linear Regression
  - Neural Networks
- Unsupervised Learning
- Ensemble Methods
- Deep Learning & Generative AI
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering
  - Regularization and Overfitting
  - Experimental Design
  - Societal Implications

What is Machine Learning?

Source: https://en.wikipedia.org/wiki/Panzanella

Things Machine Learning Isn't

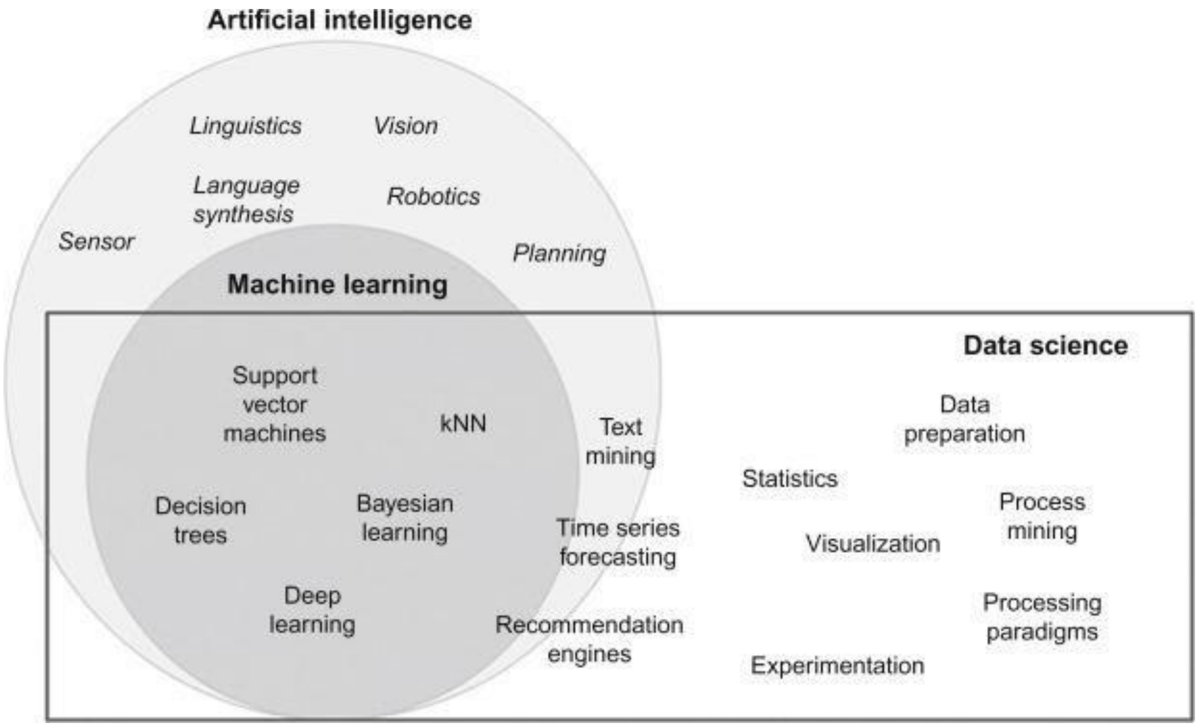
- Artificial intelligence
- Data science

Things Machine Learning Isn't

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science

Things Machine Learning Isn't

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data



Source: https://www.sciencedirect.com/topics/physics-and-astronomy/artificial-intelligence

Things Machine Learning Isn't

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral?

Defining a Machine Learning Problem (Mitchell, 97)

- A computer program **learns** if its *performance*, *P*, at some *task*, *T*, improves with *experience*, *E*.
- Three components
  - Task, T
  - Performance metric, P
  - Experience, E

Defining a Machine Learning Problem: Example

- Learning to approve loans/lines of credit

- Three components
  - Task, T

    1 Predicting risk of losing money / 2 Default

    3 Predicting amount they can pay back
  - Performance metric, P

    0-1 accuracy / ms loss / money lost / net-profit
  - Experience, E

    historical data

Problem Formulation

- Often, the same task can be formulated in more than one way.

Example: Loan applications

- creditworthiness/score (regression)
- probability of default (density estimation)
- loan decision (classification)

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Class Activity

1. Select a **task**, T
2. Identify **performance measure**, P
3. Identify **experience**, E
4. Report ideas back to rest of class

Example Tasks

- Identify objects in an image
- Translate from one human language to another
- Recognize speech
- Assess risk (e.g. in loan application)
- Make decisions (e.g. in loan application)
- Assess potential (e.g. in admission decisions)
- Categorize a complex situation (e.g. medical diagnosis)
- Predict outcome (e.g. medical prognosis, stock prices, inflation, temperature)
- Predict events (default on loans, quitting school, war)

Your Well-posed ML Problems

task, T	performance, P	experience, E
Predict bacterial behavior	Does prediction map to behavior	longitudinal data (past 3 hours)
Whether tumor is benign	Does the prediction map prognosis	Tumor imaging
Evaluate titles match content with	Click-rate (clickbait) Time-spent	human annotated articles and how they map to content

Things Machine Learning Isn't

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016

Source: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Things Machine Learning Isn't

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral

OPPORTUNITIES AND CHALLENGES IN BIG DATA

The Assumption: Big Data is Objective

It is often assumed that big data techniques are unbiased because of the scale of the data and because the techniques are implemented through algorithmic systems. However, it is a mistake to assume they are objective simply because they are data-driven.[13]

The challenges of promoting fairness and overcoming the discriminatory effects of data can be grouped into the following two categories:

1) Challenges relating to ***data used as inputs*** to an algorithm; and
2) Challenges related to ***the inner workings of the algorithm itself***.

Source: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

attributes — features; ground-truth — labels

data points / instance

	Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
patient 1	Yes	Low	Normal	No
2	No	Medium	Normal	No
	No	Low	Abnormal	Yes
	Yes	Medium	Normal	Yes
5	Yes	High	Abnormal	Yes
new patient	No	High	abnormal	Yes / No

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(<u>supervised</u>) binary classification task**

features | labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

features | labels

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **<u>classification</u> task**

features labels

data points

Family History	Resting Blood Pressure	Cholesterol	Risk
Yes	Low	Normal	Low Risk
No	Medium	Normal	Low Risk
No	Low	Abnormal	Medium Risk
Yes	Medium	Normal	High Risk
Yes	High	Abnormal	High Risk

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **regression task**

features | targets

data points

Family History	Resting Blood Pressure	Cholesterol	Medical Costs
Yes	Low	Normal	$0
No	Medium	Normal	$20
No	Low	Abnormal	$30
Yes	Medium	Normal	$100
Yes	High	Abnormal	$5000

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

$\overbrace{\text{Family History} \quad \text{Resting Blood Pressure} \quad \text{Cholesterol}}^{\text{features}}$ $\overbrace{\text{Heart Disease?}}^{\text{labels}}$

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Is this a “good” Classifier?

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

features | | | labels

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)

training dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **error rate** is the proportion of data points where the prediction is wrong

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **test error rate** is the proportion of data points in the test dataset where the prediction is wrong ($1/3$)

A Typical (Supervised) Machine Learning Routine

- Step 1 – **training**
  - Input: a labelled training dataset
  - Output: a classifier
- Step 2 – **testing**
  - Inputs: a classifier, a test dataset
  - Output: predictions for each test data point
- Step 3 – **evaluation**
  - Inputs: predictions from step 2, test dataset labels
  - Output: some measure of how good the predictions are; usually (but not always) error rate

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset

labels

data points

Heart Disease?
No
No
Yes
Yes
Yes

- This classifier completely ignores the features…

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset

$$\overbrace{\text{Heart Disease?}}^{\text{labels}}$$

data points

Heart Disease?	Predictions
No	Yes
No	Yes
Yes	Yes
Yes	Yes
Yes	Yes

- The training error rate is 2/5

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote

	Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
	Yes	Low	Normal	No
	No	Medium	Normal	No
	No	Low	Abnormal	Yes
4	Yes	Medium	Normal	Yes
	Yes	High	Abnormal	Yes
pat. 6	No	High	Abnormal	Yes
pat 7	Yes	Med	Normal	Yes

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	Predictions
Yes	Low	Normal	No	No
No	Medium	Normal	No	No
No	Low	Abnormal	Yes	Yes
Yes	Medium	Normal	Yes	Yes
Yes	High	Abnormal	Yes	Yes

- The training error rate is 0!

Is the memorizer learning?

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	Predictions
Yes	Low	Normal	No	No
No	Medium	Normal	No	No
No	Low	Abnormal	Yes	Yes
Yes	Medium	Normal	Yes	Yes
Yes	High	Abnormal	Yes	Yes

- The training error rate is 0!

data points : < features , label >

Notation

- Feature space, $\mathcal{X}$ $\vec{x}^{(i)} \in \mathcal{X}$
- Label space, $\mathcal{Y}$ $y^{(i)} \in \mathcal{Y} = \{yes, no\}$
- (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$ $y^{(i)} = c^*(\vec{x}^{(i)})$
- Training dataset: $\mathcal{D} = \{\underbrace{<\boldsymbol{x}^{(1)}, y^{(1)}>}_{\text{row 1}}, \ldots, \underbrace{<\boldsymbol{x}^{(N)}, y^{(N)}>}_{\text{row 5}}\}$
- Data point:
  $<\boldsymbol{x}^{(i)}, y^{(i)}> = <x_1^{(i)}, x_2^{(i)}, \ldots, x_D^{(i)}, y = c^*(\boldsymbol{x})>$
- Classifier, $h : \mathcal{X} \rightarrow \mathcal{Y}$ $\hat{y} = h(\vec{x}) =$
- Goal: find a classifier, h, that best approximates c^*

Notation

Performance

- Loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Defines how "bad" predictions, $\hat{y} = h(\vec{x})$ are
  - compared to the true labels, $y = c^*(\boldsymbol{x})$
- Common choices
  - Binary or 0-1 loss (for classification):

$$\ell(y, \hat{y}) = \mathbb{1}[y \neq \hat{y}] \rightarrow \text{indicator}$$

$\mathbb{1}(0,1) = 1$

$\mathbb{1}(0,0) = 0$

  - Squared loss (for regression):

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

$\ell(0,1) = 1$

$\ell(0,10) = 100$

- Error rate:

$$Err(h, D) = \frac{1}{N} \sum_{i=1} \ell\left(y^{(i)}, \hat{y}^{(i)}\right)$$

Notation - Practice

x_1 Family History	x_2 Resting Blood Pressure	x_3 Cholesterol	y Heart Disease?	$\hat{y}$ Prediction
Yes	Low	Normal	No	Yes
No	Medium	Normal	No	No
No	Low	Abnormal	Yes	No
Yes	Medium	Normal	Yes	Yes
Yes	High	Abnormal	Yes	Yes

$x = (\text{Family history}, \text{RBP}, \text{Chol})$

$y = \{\text{yes}, \text{no}\}$

$N = 5$

$D = 3$

$h(\vec{x}) = \text{'yes'}$

$h'(\vec{x}) = \mathbb{1}[x_1 = \text{Yes}]$

$\frac{1}{5} \sum \ell(y^i, \hat{y}^i) = 0.4$

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote
- The memorizer (typically) does not **generalize** well, i.e., it does not perform well on unseen data points
- In some sense, good generalization, i.e., the ability to make accurate predictions given a small training dataset, is the whole point of machine learning!

Learning Goals

- You should be able to

1. Formulate a well-posed learning problem for a real-world task by identifying the task, performance measure, and training experience
2. Describe the supervised learning paradigm in terms of the type of data needed, the form of prediction, and the structure of the output prediction
3. Explain the difference between memorization and generalization

Logistics: Course Syllabus

https://www.cs.cmu.edu/~10701-f25/

This whole website is **required** reading.