CS 559 Machine Learning

Lecture 1: Introduction and Overview

Ping Wang
Department of Computer Science
Stevens Institute of Technology

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Today's Lecture

- Basic course information
- Goal of this course
- Grading policy
- Course overview
- Overview of probability

Course Information

- Meeting:
  - **Time**: Thursday, 3:30PM–6:00PM, 2023 Spring
  - **Location**: Howe 102
- Canvas:
  - Announcements, Assignments, Discussions.
  - Login to myStevens
  - Course Email: mlfa.stevens@gmail.com
  - Introduce yourself on Canvas discussions

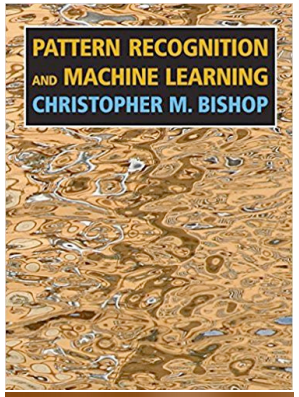
Instructor and TA Information

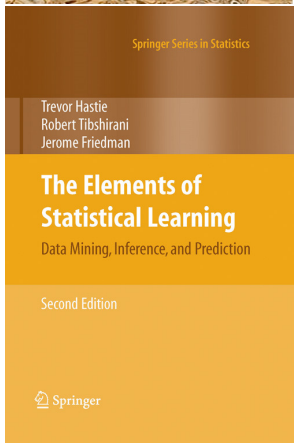
- Instructor: Ping Wang
  - **Office hours**: Monday, 3:00PM-4:00PM via Zoom
  - **Zoom link**: https://stevens.zoom.us/j/97789938972
  - **Email**: ping.wang@stevens.edu

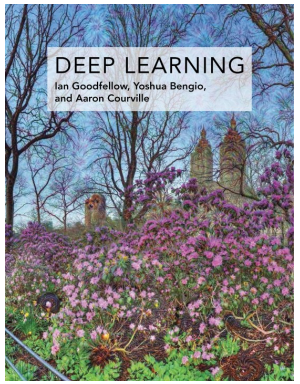
- Teaching Assistant:
  - Xinming Yang
    - **Office hours**: Monday, 1:00PM-2:00PM via Zoom
    - **Zoom link**: https://stevens.zoom.us/j/91416409329
    - **Email**: xyang70@stevens.edu
  - Ayush Ashutosh Panigrahi: TA information will be provided soon.

Textbooks

- Pattern Recognition and Machine Learning
  - Bishop, Christopher M.
  - http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf
- The Elements of Statistical Learning
  - Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome
  - https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf
- Deep Learning
  - Ian Goodfellow and Yoshua Bengio and Aaron Courville
  - https://www.deeplearningbook.org/







Prerequisites

- Prerequisite Course
  - MA 222 Probability Theory, or CS 556 Mathematical Foundations of Machine Learning, or Equivalent.
- Programming
  - Python
- Linear Algebra
  - Vector, matrix, projection, eigenvalues/vector ...
- Probability, Statistics and Optimization
  - Distributions, expectation/variance, objective function, ...

Goal of This Course

- Be more proficient at math and programming.
- Be able to recognize when and how a new problem can be solved with an existing technique.
- Be able to implement general ML techniques for a variety of problem types.

Grading Policy

- Distribution:
  - Homework: 35%;
  - Midterm Exam: 30%
  - Final Exam: 30%
  - Participation: 5%

Grading

- **Distribution of (0-100):**
  - A (90-100)
  - A- (85-90)
  - B+ (80-85)
  - B (75-80)
  - B- (70-75)
  - C+ (65-70)
  - C (60-65)
  - F (<60)

Final Grade Distribution in Fall 2022

45 students in total:

- A: 20
- A-: 7
- B+: 4
- B: 6
- B-: 1
- C+: 3
- C: 3
- F: 1

Homework

- **Goal:** test your ability to understand knowledge of ML
  - Five homework, HW5 will be optional.
  - Only four homework with the highest score will be considered in the final grading.
  - Mix of written and programming problems.
- **Submission:**
  - Upload your e-copies (PDF file) on **Canvas**
  - **Jupyter Notebook** is recommended to use for your programming assignments (with code and output after execution). Make sure to include detailed comments and analysis in your code.

Late Policy

- All the homework assignments must be submitted on Canvas before 11:59 PM on the due date.
- Any late submission within 24 hours will be penalized 10%.
- Any late submission within 24-48 hours will be penalized 20%.
- Any late submission within 48-72 hours will be penalized 40%.
- The 72 hours after the deadline will be the hard deadline for each assignment. Assignments submitted after this hard deadline will not be graded and get no points for the assignment.

Honor Code

- Collaborate and discuss together but write code and reports independently.
- Do not look at classmates' writeup or code
- Do not share writeup/code (online and physical)
- Debugging: only look at input-output behavior
- Detect plagiarism.
- If same/similar solutions or codes are submitted, you will not get any points.

Emergency Resources

Stevens Campus Police
Kidde Building, Ground Floor
201-216-3911 (Emergency Line, 24/7)
201-216-5105 (Non-Emergency Line, 24/7)

National Suicide Prevention Lifeline
1-800-273-8255

Crisis Textline
Text HOME to 741-741 (24/7)

Please save these numbers in your phone.

Emergency Communication

Sign up or update your information in Stevens Alerts, the emergency alert system used at Stevens. Be in the know about snow days and campus emergencies.

Be sure to provide:

- Your Stevens email
- Your personal email
- Your cell phone number

Go to MyStevens; click on 'Stevens Alerts' to register.

Report a Concern

If you have a concern about another student that is ***NOT time sensitive***, use the '*Report a Concern*' link on MyStevens or email care@stevens.edu to inform a team of professionals who can assist.

Wellness Resources

Counseling and Psychological Services (CAPS)
Wellness Center, 2nd Floor; 201-216-5177
Free personal and group counseling.
Call for an appointment.
Counselors are available 24/7.

Student Health Services
Wellness Center, 1st Floor; 201-216-5678
Call in advance to be seen by a clinician.
A nurse is available 24/7.

Disability Services
Wellness Center, 2nd Floor
Assists students with disabilities to fully participate in campus services and programs with equal access.

Course Overview

Acknowledgements

- Course Textbooks
- Yue Ning and Tian Han’s lectures at Stevens
- Chandan K. Reddy’s lectures at Virginia Tech
- Other online courses and materials

What is Machine Learning?

- The term “Machine Learning” was first coined in 1959 by Arthur Samuel from IBM. It is a branch of Artificial Intelligence (AI).
- Goal: focused on design and development of algorithm to improve their performance
- Input: empirical data, such as that from sensors or databases
- Output: ***patterns*** or ***predictions*** thought to be features of the underlying mechanism that generated the data.
- Learner (the algorithm): takes advantage of ***data*** to capture ***characteristics of interest*** about their unknown underlying probability distribution.

AI and Machine Learning

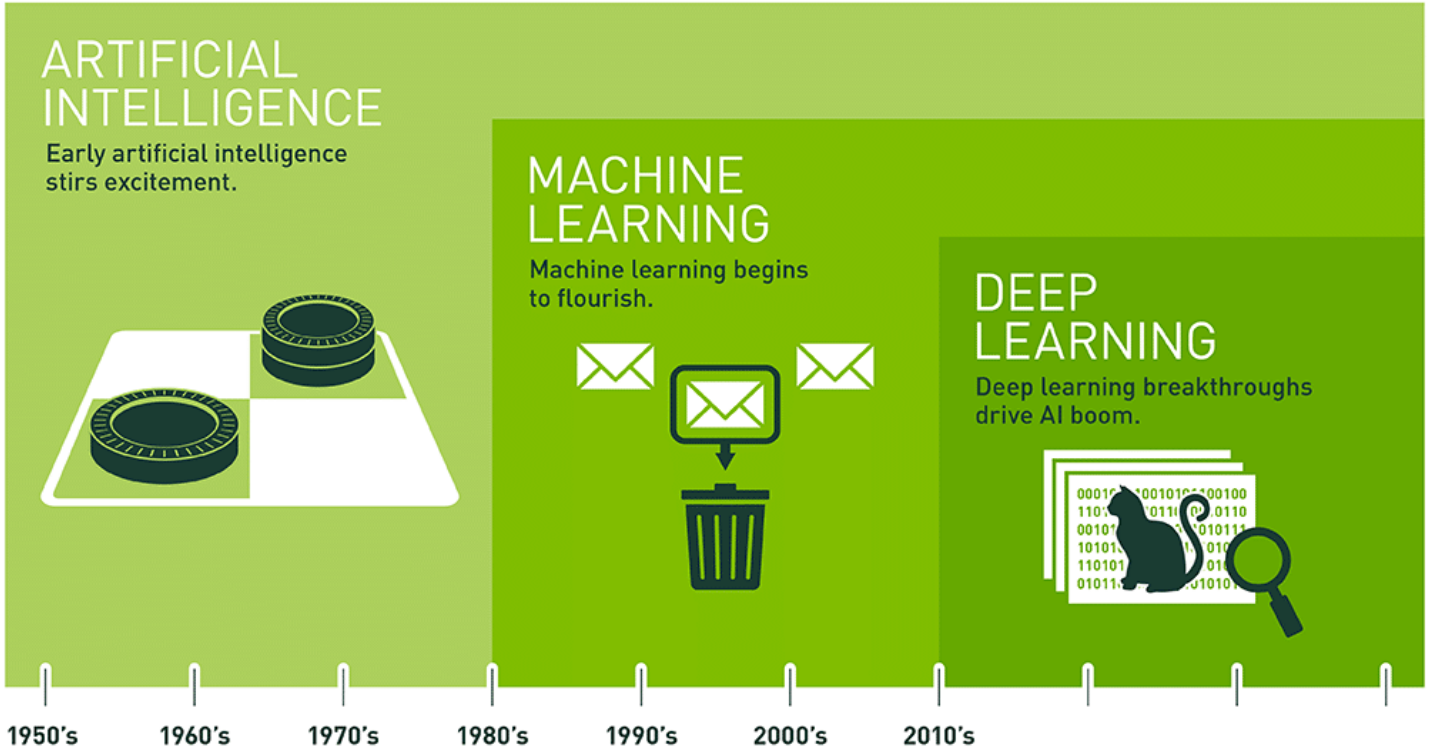
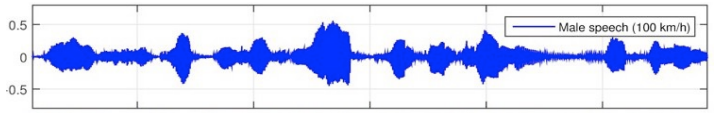


Figure: Nvidia blog about "What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?".

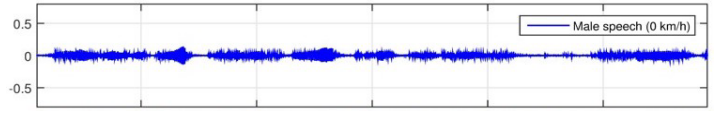
Example of ML: Speech Recognition

- Given an audio waveform, robustly extract & recognize any spoken words.

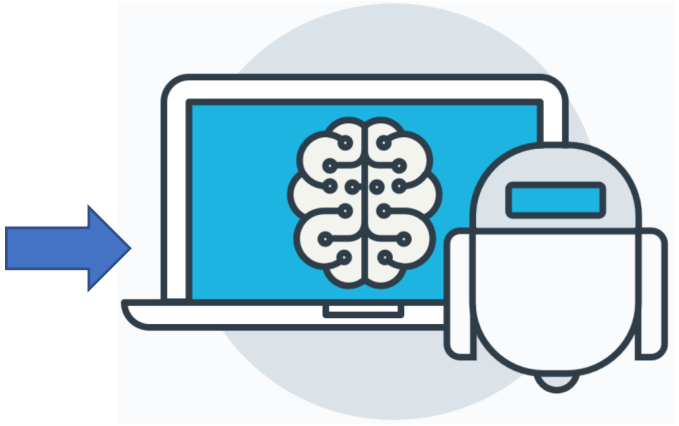
"How is it going"



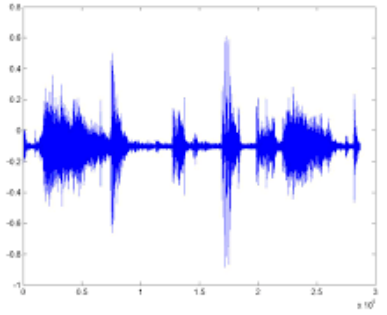
"Glad to see you"



Training audio data



Program for learning



"Nice to meet you"

Testing audio data

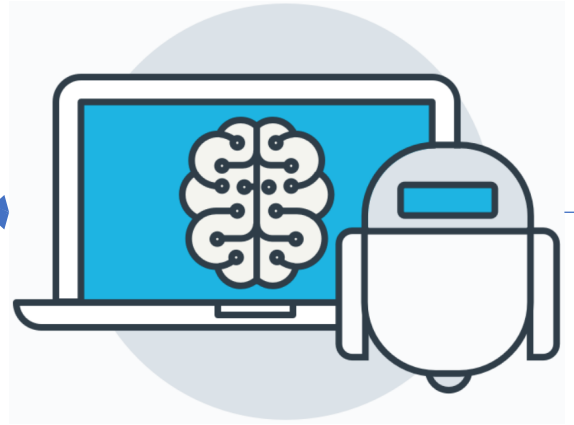
Example of ML: Image Recognition

Monkey

Dog

Cat

Training image data

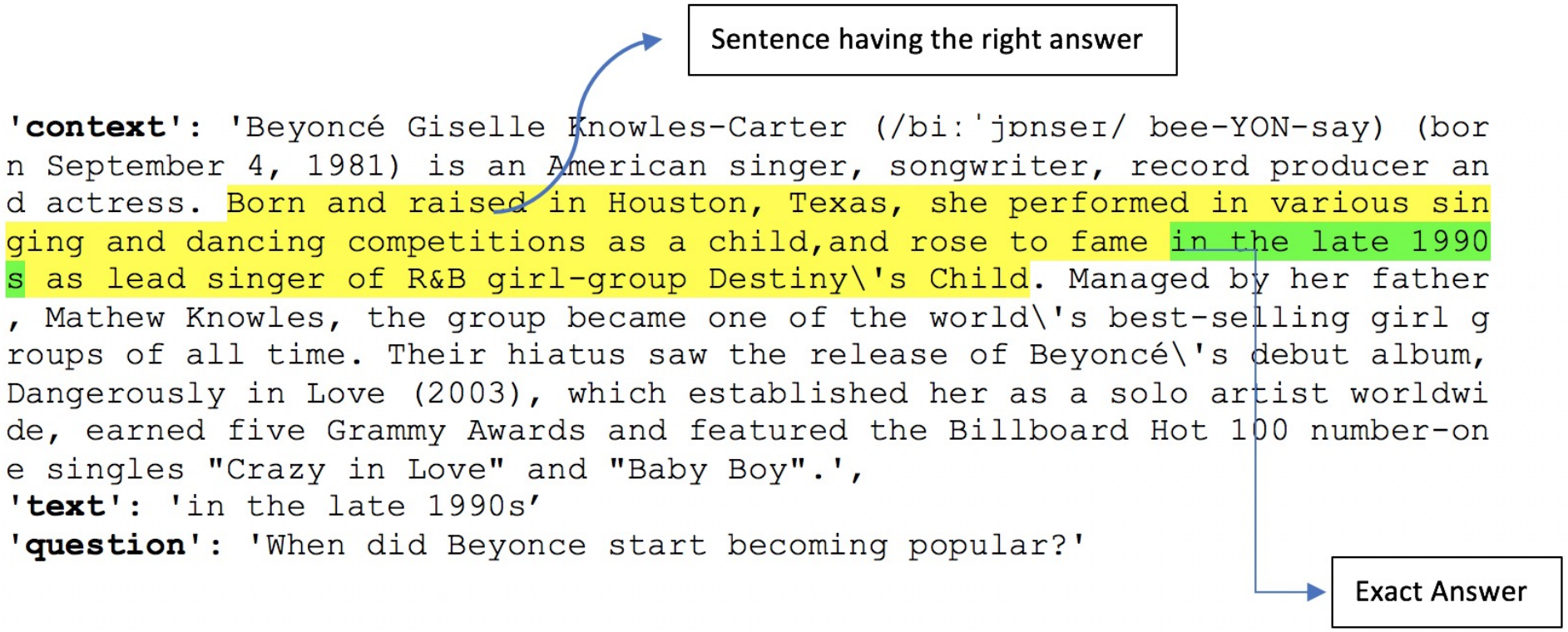


Cat

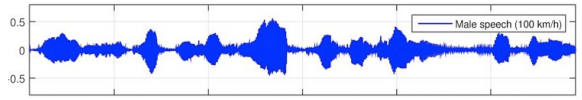
Testing image data

Example of ML: Question Answering

- Given a textual snippet and a question, accurately identify the text spans as the answer to the question.



Machine Learning ≈ Looking for a function

- Speech recognition: $f($ 

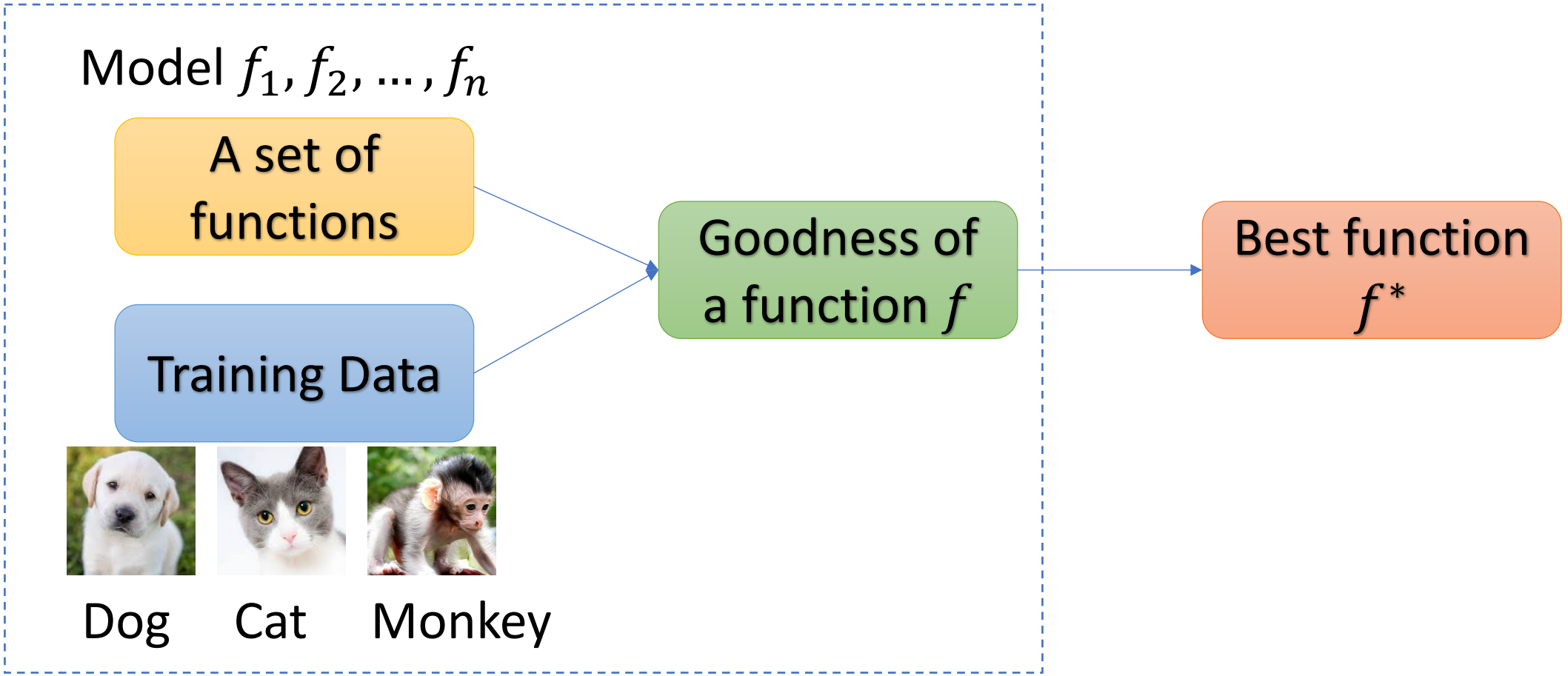
$)$ → "Nice to meet you"

- Image recognition: $f($  $)$ → "Cat"

- Question answering: $f\left(\begin{matrix}\text{"when did Beyonce start becoming popular?"}\\ a\ text\ snippet\end{matrix}\right)$

→ "in the late 1990s"

Training Framework



Testing Framework

Many Scientific Fields

- **Computer science**: Artificial intelligence, computer vision, information retrieval,...
- **Statistics:** Inference from data, probabilistic models, learning theory, ...
- **Mathematics:** Optimization theory, numerical methods, tools for theory,...
- **Engineering:** Signal processing, system identification, robotics, control, …
- **Economics:** decision theory, operations research, econometrics, ...
- **Biomedical:** Health informatics, Precision medicine, Epidemic control, …
- **Chemistry:** Molecular drug discovery, Chemical engineering, …

Along with many others!

Components in ML Algorithms

- Every machine learning algorithm has three components:
  - Representation / Model Class
  - Evaluation / Objective Function
  - Optimization

Representation / Model Class

- Decision trees
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles

Evaluation / Objective Function

- Accuracy
- Precision and recall
- F1 score
- Squared error
- Likelihood
- Posterior probability
- Margin
- Entropy
- K-L divergence

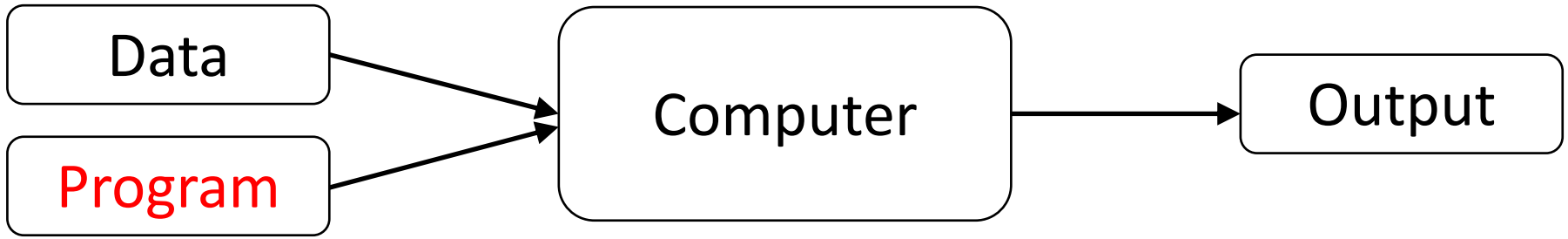
Optimization

- Discrete optimization
  - Minimal Spanning Tree
  - Shortest Path

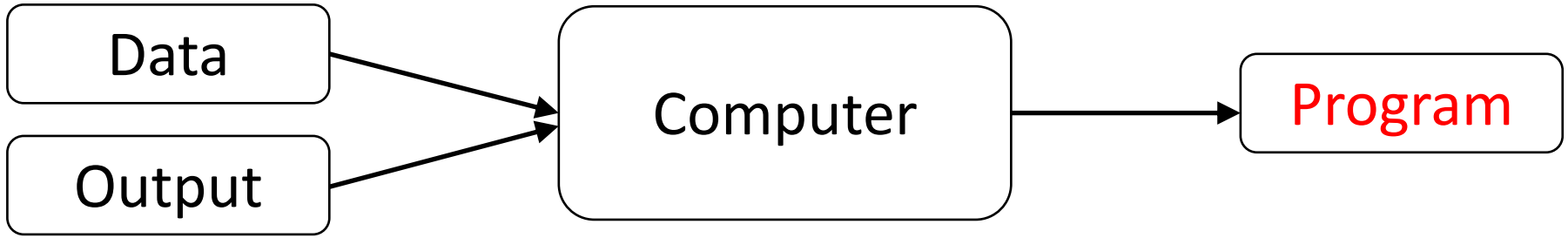
- Continuous Optimization
  - Gradient Descent
  - Linear Programming

ML vs Traditional Approach

- Traditional Approach



- Machine Learning



Challenges in Machine Learning

- Data:
  - Lack of training data
  - Nonrepresentative training data
  - Poor quality of data: noise, errors, outliers, …
  - Irrelevant features: need good feature engineering
- Algorithm:
  - Overfitting
  - Underfitting
  - Poor generalization
  - Biased prediction

Types of Machine Learning

- Supervised Learning
  - Training data include desired outputs
  - Test data only have features, must predict outputs

- Unsupervised learning
  - Training data do not include desired outputs

Supervised Learning

- Sample data comprises input features along with the corresponding target values(labeled data).
- Supervised learning uses the given labeled data to find a model (hypothesis) that predicts the target values for previously unseen data.

Supervised Learning: Regression

- Each data sample is provided with one or more feature variables, which are associated with a scalar continuous output variable.
- Example: given the gender and height information, predict the weight.

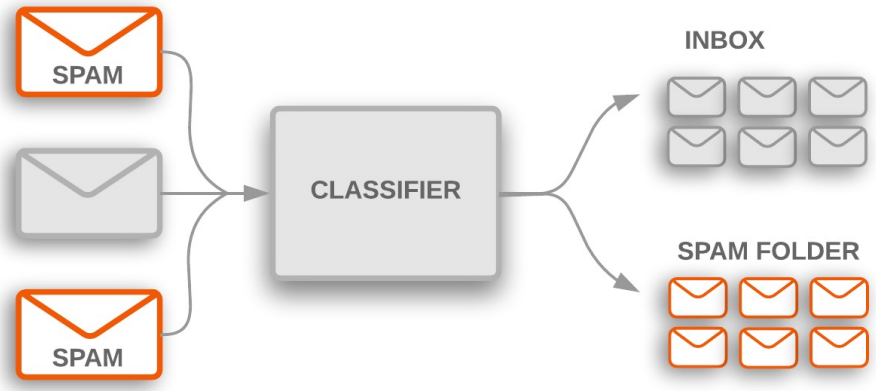
Data sample	X1: Gender	X2: Height (inches)	Y: Weight (pounds)
1	Male	73	190
2	Male	67	165
3	Female	65	130
	Features/Attributes		**Output**

Supervised Learning: Classification

- Each data sample is labeled as belonging to some class. There is no order among classes.

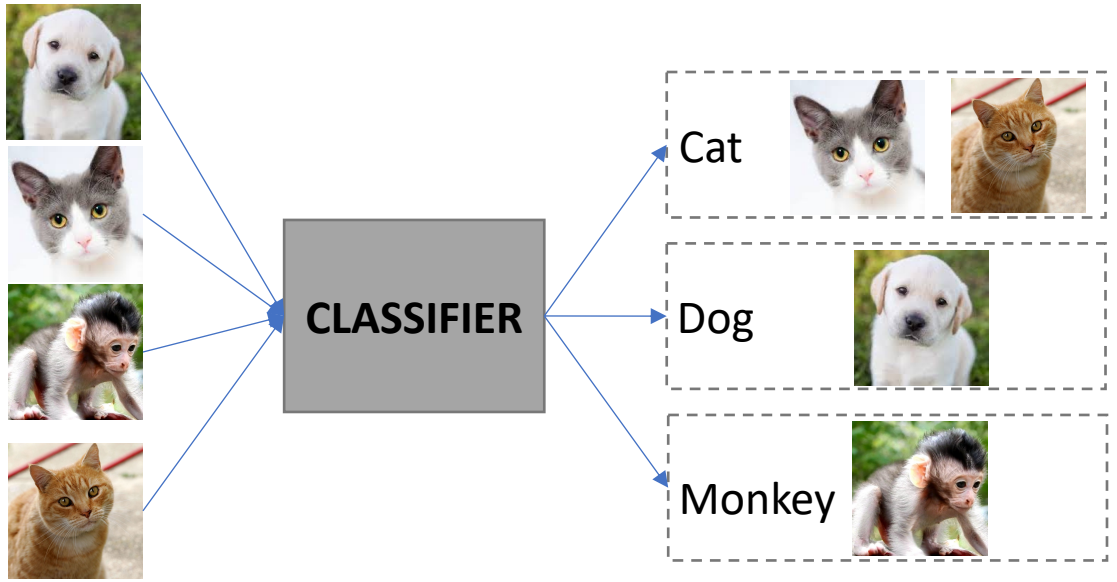
Binary classification

Output: Yes (1) or No (0)



Multi-class classification

Output: one of three or more classes



Special Cases of Supervised Learning

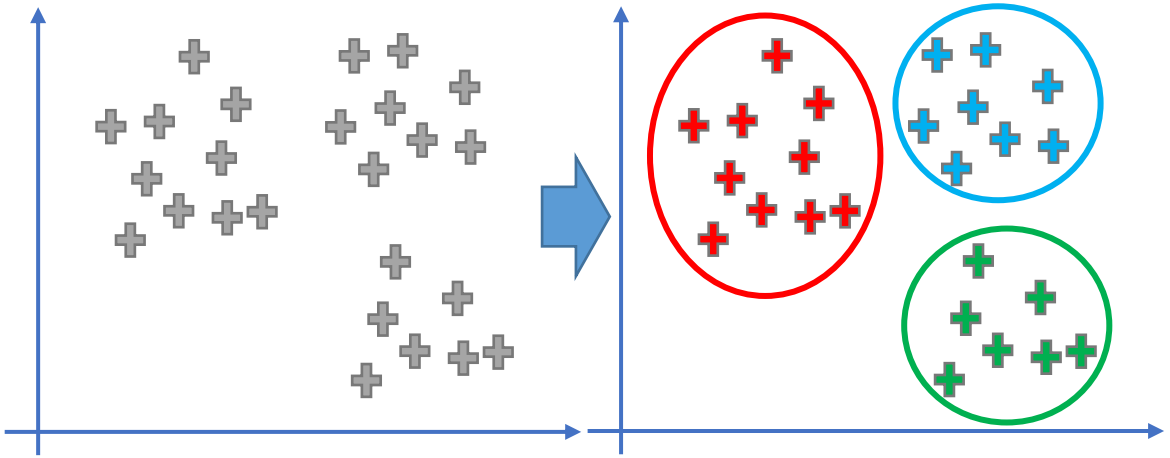
- Semi-supervised learning:
  - Incomplete training data, some of the target outputs missing
  - Besides labeled examples, it exploits unlabeled ones to improve the generalization ability of the classifier.
- Active learning
  - A limited set of training labels based on budget
  - Have to optimize the choice of objects to acquire labels
  - When used interactively, these can be presented to the user for labeling
- Many others …

Unsupervised Learning

- The given data consists of input features without any corresponding target values.
- The goal is to discover groups of similar examples within the data (clustering), or to determine the distribution of data within the input space (density estimation).

Unsupervised Learning

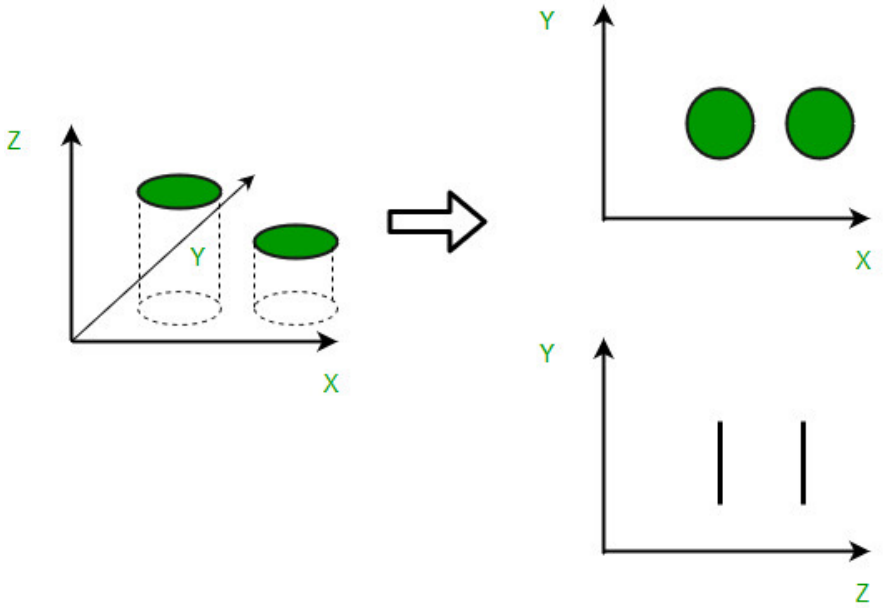
➢Clustering



Before clustering

After clustering

➢Dimension reduction



Before transformation

After transformation

Self-supervised Learning

- **Motivation**: leverage the unlabeled data pool to learn useful representation of the data using self-supervision, which can reduce the data labelling cost.
- **Idea**: create some auxiliary supervised task from the unlabeled data.
- Popular in natural language processing (NLP) and computer vision.

Self-supervised Learning in Computer Vision

➢Small distortion (coloring, rotation, scaling) on an image does not modify its original semantic meaning or geometric forms, thus the learned features are expected to be invariant to distortion.

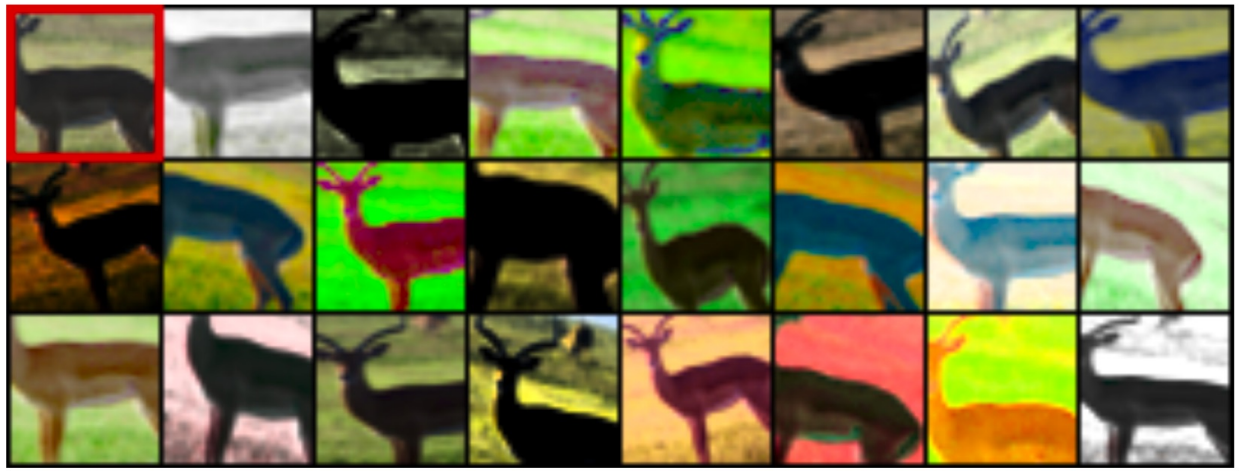
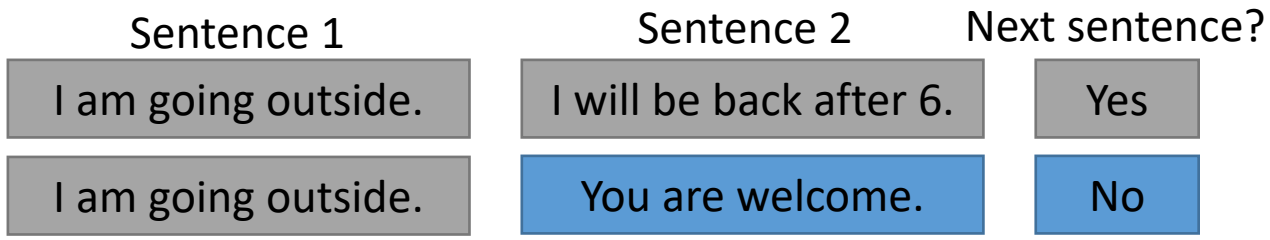


Image source: Dosovitskiy et al., Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks, 2015.

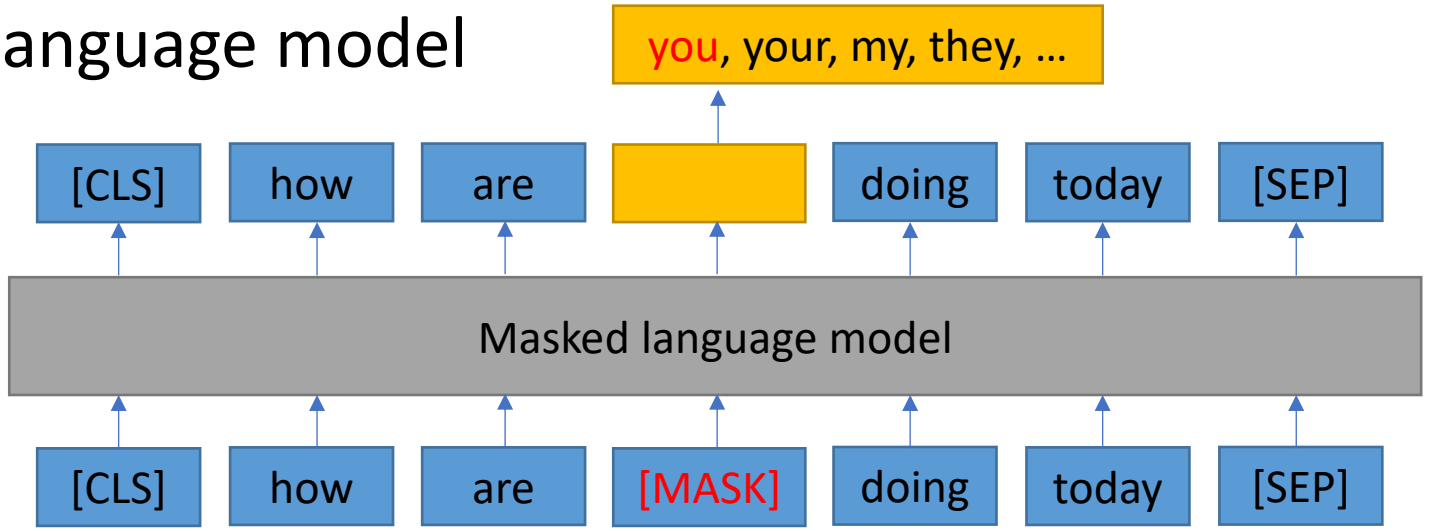
- Original patch of a cute deer in the top left corner.
- Random transformations are applied, resulting in a variety of distorted patches.
- All of them should be classified into the same class.

Self-supervised Learning in NLP

- Next sentence prediction



- Masked language model



Devlin J, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

Topics Covered in This Course (1)

- **Decision Theory** - Explain Bayesian decision theory, the likelihood ratio, and minimum risk classification.
- **Maximum Likelihood Estimation** - Implement Maximum Likelihood Estimation for Logistic Regression.
- **Dimensionality Reduction** - Apply dimensionality reduction using Principal Component Analysis.
- **Linear Discriminant Functions** - Implement classifiers using linear discriminant functions and Fisher Linear Discriminant Analysis.

Topics Covered in This Course (2)

- **Non-parametric Learning** - Implement k-nearest neighbors, and perform non-parametric classification.
- **Clustering** – Implement k-means clustering, and perform EM for Gaussian mixtures.
- **Support Vector Machines** - Explain the advantages of Support Vector Machines and margin maximization.
- **Boosting** - Explain boosting and decision tree models.
- **Neural Networks** – Implement backpropagation for basic neural networks.

Overview of Probability

Discrete Random Variable

- Discrete random variable is a variable whose value is obtained by counting.
  - Example: number of students in the class, students' grade level
- Discrete random variables either have a finite or countable number of states (possible values).
- Event: a set of outcomes of an experiment. It is a possible combination of different values of a random variable.

Discrete Random Variable

- Probability mass function $P(X = x)$: A function which tells us how likely each possible outcome is.

$$P(X = x) = P_X(x) = P(x)$$

$$P(x) \geq 0 \; for \; each \; x$$

$$\sum_{x \in \Omega} P(x) = 1$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x)$$

- Example: Bernoulli, Binomial, Multinomial, Poisson

Conditional Probability

- Conditional probability: Recalculated probability of event A, given that another event B has already occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\Rightarrow P(A \cap B) = P(A|B)P(B)$$

- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Expectation and Variance

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments: expectation of power of X: $M_k = E(X^k)$
- Variance: expectation of the squared deviation from the mean of X

$$\begin{aligned} Var(X) &= E\big(X - E(X)\big)^2 \\ &= E(X^2) - (E(X))^2 \\ &= M_2 - M_1^2 \end{aligned}$$

- Standard deviation: Square root of variance.

Bivariate Distributions

- Joint distribution: $P(X = x; Y = y)$, a list of all probabilities of all possible pairs of observations.
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $P(X|Y)$ specifies a "lookup-table" of all possible $P(X = x|Y = y)$.

Conditioning and marginalization come up in Bayesian inference ALL the time: Condition on what you observe. Marginalize out the uncertainty.

Expectation and Covariance of Multivariate Distributions

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Covariance is the expected value of the product of deviation:

$$Cov(X,Y) = E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)\right)$$
$$= E(XY) - E(X)E(Y)$$

$$Var(X) = Cov(X,X)$$

- Aside: One common way to construct bivariate random variables is to have a random variable, and the parameter of its distribution is another random variable.

Independence of Random Variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y .
- If X and Y are independent, $Cov(X, Y) = 0$.
- Aside: Mutual information is a measure of how “non-independent” two random variables are.

Multivariate Distributions

- X, x are vector-valued.
- Mean: $E(X) = \sum_x xP(x)$
- Covariance matrix:

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$$
$$Cov(X) = E(XX^T) - E(X)E(X)^T$$

- Conditional and marginal distributions: Can define and calculate any (multi or single-dimensional) marginals or conditional distributions we need: $P(X_1), P(X_1, X_2), P(X_1, X_2, X_3|X_4)$, etc..

Example (from Bishop. [2006])

- Assuming, we know that:

$$P(B = r) = 4/10$$
$$P(B = b) = 6/10$$

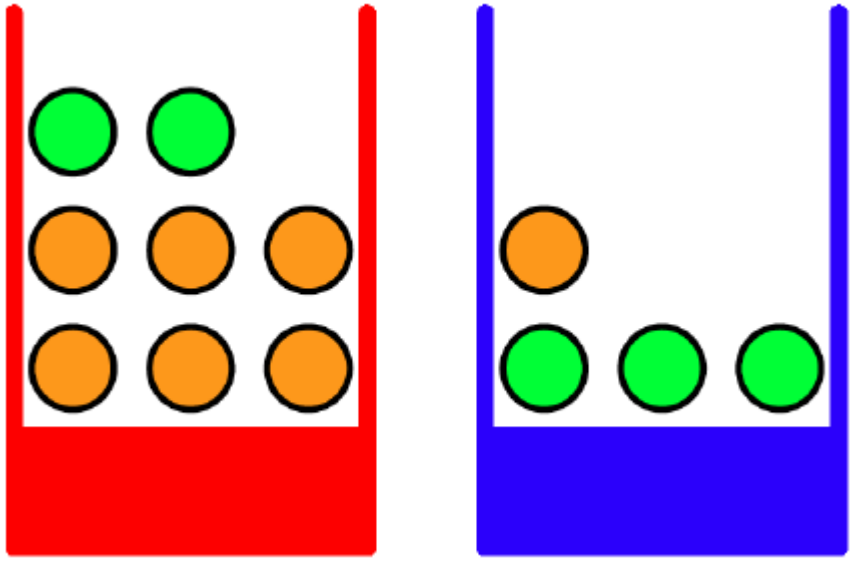
- The probability of selecting a fruit from a given box is:

P(F = a|B = r) = 1/4

P(F = o|B = r) = 3/4

P(F = a|B = b) = 3/4

P(F = o|B = b) = 1/4



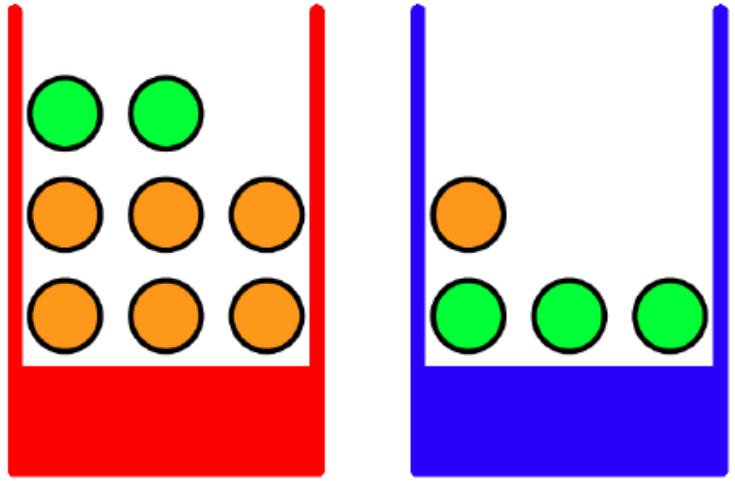
Example (from Bishop. [2006])

- What is the probability of choosing an apple?
- Conditional probability:

$$P(F = a) = P(F = a|B = r)P(B = r) + P(F = a|B = b)P(B = b) = \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20}$$

- Thus, the probability of choosing an orange is

$$P(F = o) = 1 - \frac{11}{20} = \frac{9}{20}$$

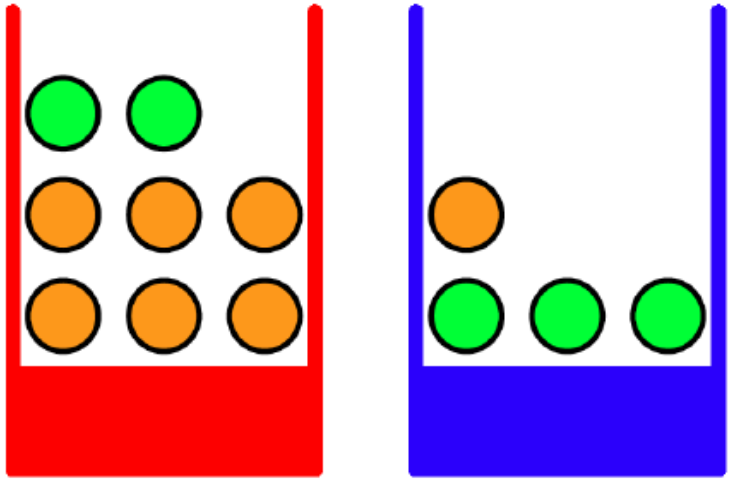


Example (from Bishop. [2006])

- We are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.
- Using Bayes' theorem,

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)}$$

$$= \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = \frac{2}{3}$$



Prior vs. Posterior

- Prior Probability
  - If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.
- Posterior Probability
  - Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.

Bayesian Probabilities

- Bayesian view: we can adopt a similar approach for parameter inference. The probabilities provide a quantification of uncertainty for the model parameters.
- Before observing the data, the assumptions about w are captured in the form of a prior probability distribution $P(w)$.
- The effect of the observed data $D = \{(x_1, y_1), \dots (x_N, y_N)\}$ is expressed by the conditional probability $P(D|w)$.
- Bayes' theorem:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}$$

- Bayes' theorem in words:

$$posterior \propto likelihood \times prior$$

Continuous Random Variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx$$

$$P(a < X < a + dx) \approx p(a) \cdot dx$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.

Continuous Random Variables

- More generally: If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1$$

Mean, Variance and Conditionals

- Mean: $E(X) = \int_x \; x \cdot p(x)dx$
- Variance: $Var(X) = E(X^2) - (E(X))^2$
- Example: Uniform
- If X has pdf $p(x)$, then $X|(X \in A)$ when $P(A) > 0$ has pdf

$$P_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx}$$

Bivariate Continuous Distributions

- $p_{X,Y}(x, y)$, joint probability density function of X and Y
- $\int_x \int_y p(x, y)dxdy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$
- Independence: X and Y are independent if $p_{X,Y(x,y)} = p_X(x)p_Y(y)$

The Univariate Gaussian

- Probability density function:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2)$$

- Easy to validate:

$$\int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1$$

- Expectation:

$$E(x) = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)xdx = \mu$$

- Variance:

$$Var(x) = E(x^2) - (E(x))^2 = \sigma^2$$

Products of Gaussian pdfs

- Suppose $p_1(x) = N(x, \mu_1, 1/\beta_1)$ and $p_2(x) = \mathcal{N}(x, \mu_2, 1/\beta_2)$, then

$$p_1(x)p_2(x) \propto \mathcal{N}(x, \mu, 1/\beta)$$

$$\beta = \beta_1 + \beta_2, \mu = \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2)$$

- In general:

$$p_1(x)p_2(x)\cdots p_n(x) \propto \mathcal{N}(x, \mu, 1/\beta)$$

$$\beta = \sum_n \beta_n \,, \mu = \frac{1}{\beta}\sum_n \mu_n\beta_n$$

- This is also true for multivariate Gaussians!

Maximum Likelihood (ML) Estimation (Gaussian)

- Assuming data points are independent and identically distributed (i.i.d.) from a Gaussian distribution, to determine the parameters from the dataset.
- The probability of the data set given μ and σ^2 (the likelihood function):

$$P(x|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

- The corresponding Log-likelihood function:

$$logP(x|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}log\sigma^2 - \frac{N}{2}\log(2\pi)$$

- Maximizing Log-likelihood with respect to μ and σ^2 by setting their derivatives to 0:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n\,,\ \ \sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

Maximum Likelihood (ML) Estimation

- **Four steps** of Maximum Likelihood (ML) Estimation:
  1) Get the likelihood function based on the given distribution.
  2) Get the corresponding log-likelihood function.
  3) Get the derivative of the log-likelihood function with respect to each parameter.
  4) Set the derivative to 0 and get the estimations.

More readings about MLE: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading10b.pdf

Maximum Likelihood Estimation (Gaussian)

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n\,,\ \ \sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

- The ML solutions μ_{ML} and σ_{ML}^2 are functions of the data set values $x_1, \ldots, x_N$. The expectations of these quantities w.r.t the data set values:

$$E(\mu_{ML}) = \mu$$

$$E(\sigma_{ML}^2) = \left(\frac{N-1}{N}\right)\sigma^2$$

- On average, the ML estimator obtains correct means but underestimate the true variance by a factor $\frac{N-1}{N}$.

Maximum Posterior (MAP) Estimator

- Given N input values $x = (x_1, \dots, x_N)^T$ and their corresponding target values $y = (y_1, \dots, y_N)^T$. We can express our uncertainty over the values of the target variables using a Gaussian distribution with a mean $f(x, w)$:

$$p(y|x, w, \beta) = \mathcal{N}(y|f(x, w), \beta^{-1})$$

- Using our training data to determine the unknown parameters w, β by maximum likelihood:

$$p(y|x, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|f(x_n, w), \beta^{-1})$$

- Log Likelihood:

$$\ln p(y|x, w, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(f(x_n, w) - y_n)^2 + \frac{N}{2} ln\beta - \frac{N}{2}\ln(2\pi)$$

Maximum Posterior (MAP) Estimator

- A more Bayesian approach by introducing a prior distribution (Gaussian) for polynomial coefficients w:

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left(-\frac{\alpha}{2} w^T w\right)$$

- Using Bayes' theorem, the posterior distribution for w:

$$p(w|x, y, \alpha, \beta) \propto p(y|x, w, \beta) p(w|\alpha)$$

- Taking the negative logarithm, Maximum posterior (MAP) is equivalent to minimizing the regularized sum-of-squares error function:

$$\frac{\beta}{2} \sum_{n=1}^{N} \{f(x_n, w) - y_n\}^2 + \frac{\alpha}{2} w^T w$$

Bayesian Probabilities

- A key issue in pattern recognition is uncertainty. It is due to incomplete and/or ambiguous information, i.e. finite and noisy data.
- Probability theory and decision theory provide the tools to make optimal predictions given the limited available information.
- In particular, the Bayesian interpretation of probability allows to quantify uncertainty, and make precise revisions of uncertainty in light of new evidence.

Summary of Today's Lecture

- Make sure to login to Canvas to check the basic course information, check the homework assignments and participant in discussions.
- Introduce yourself on Canvas
- Course overview
- Overview of probability