

Lecture 1: Introduction

André Martins, Francisco Melo, Mário Figueiredo



Deep Learning Course, Winter 2023-2024

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Why Did Deep Learning Become Mainstream?

Lots of recent breakthroughs:

- Object recognition
- Speech and language processing (Transformers, BERT, GPT-3)
- Machine translation
- Chatbots and dialog systems
- Self-driving cars
- Solving games (Atari, Go, StarCraft II)
- Protein design (AlphaFold)

No signs of slowing down...

“But It’s Non-Convex”

For many years (1990–2010), NNs were not popular in machine learning, because they were hard to learn.

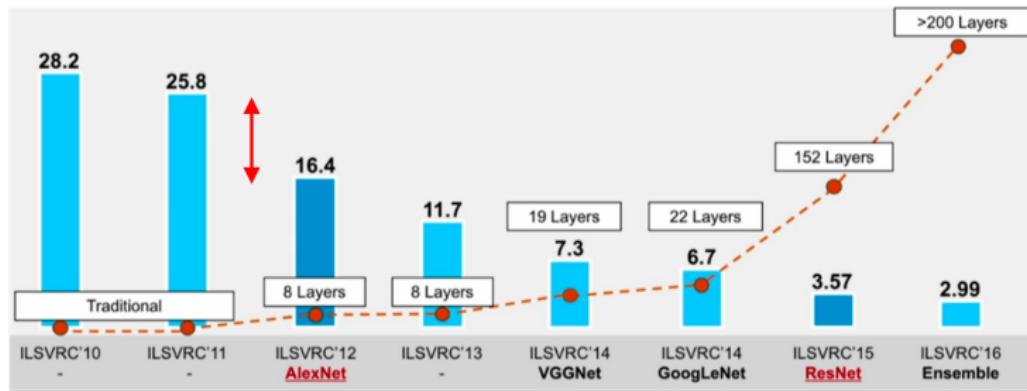
Why does gradient-based optimization work at all in NNs despite the non-convexity?

One **possible**, **partial** answer (this is an open research topic)

- there are generally many hidden units
- there are many ways a neural net can approximate the desired input-output relationship
- we only need to find one

One turning point: AlexNet

- Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton; 2012
- ImageNet: Large Scale Visual Recognition Challenge (14 million images, 20000 categories)



- Large CNN, much deeper than anything else at the time
- Used parallel processing (one of the first uses of GPUs in NNs)
- Convinced many people that deep learning would change the field.

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Today's Roadmap

- Linear Algebra (only skim over)
- Probability Refresher (only skim over)
- Optimization
- Introduction to Machine Learning
 - Supervised, Unsupervised, Reinforcement Learning
 - Classification and Regression

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m \times n}$ is a **matrix** with m rows and n columns.

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}.$$

- $x \in \mathbb{R}^n$ is a **vector** with n components,

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

- A **(column) vector** is a matrix with n rows and 1 column.
- A matrix with 1 row and n columns is called a **row vector**.

Matrix Transpose and Matrix Products

- Given matrix $A \in \mathbb{R}^{m \times n}$, its **transpose** A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- Matrix A is **symmetric** if $A^T = A$.
- Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their **product** is

$$C = AB \in \mathbb{R}^{m \times p} \text{ where } C_{i,j} = \sum_{k=1}^n A_{i,k} B_{k,j}$$

- Inner product** between vectors $x, y \in \mathbb{R}^n$:

$$\langle x, y \rangle = x^T y = y^T x = \sum_{i=1}^n x_i y_i$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^T \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = [x_1, \dots, x_n] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i$$

Outer and Hadamard Products

- Outer product between vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$:

$$x y^T \in \mathbb{R}^{n \times m}, \quad \text{where } (x y^T)_{i,j} = x_i y_j$$

$$x y^T = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} [y_1, \dots, y_m] = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_m \\ \vdots & \ddots & \vdots \\ x_n y_1 & \cdots & x_n y_m \end{bmatrix}$$

- Hadamard/Schur product between vectors $x, y \in \mathbb{R}^n$: $(x \odot y)_i = x_i y_i$,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \odot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 \\ \vdots \\ x_n y_n \end{bmatrix}$$

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their **product** is

$$C = AB \in \mathbb{R}^{m \times p} \text{ where } C_{i,j} = \sum_{k=1}^n A_{i,k} B_{k,j}$$

- Matrix product is **associative**: $(AB)C = A(BC)$.
- In general, matrix product is not **commutative**: $AB \neq BA$.
- Transpose of product: $(AB)^T = B^T A^T$.
- Transpose of sum: $(A + B)^T = A^T + B^T$.

Norms

- The **norm** of a vector is (informally) its “magnitude.” Euclidean norm:

$$\|x\|_2 = \sqrt{\langle x, x \rangle} = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}.$$

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}.$$

- Notable case: the ℓ_1 norm, $\|x\|_1 = \sum_i |x_i|$.
- Notable case: the ℓ_∞ norm, $\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}$.

Special Matrices

- The **identity matrix** $I \in \mathbb{R}^{n \times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \quad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A I = I A = A$.
- Diagonal matrix: $A \in \mathbb{R}^{n \times n}$ is diagonal if $(i \neq j) \Rightarrow A_{i,j} = 0$.
- Upper triangular matrix: $(j < i) \Rightarrow A_{i,j} = 0$.
- Lower triangular matrix: $(j > i) \Rightarrow A_{i,j} = 0$.

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an **eigenvector** of matrix $A \in \mathbb{R}^{n \times n}$ if

$$Ax = \lambda x,$$

where $\lambda \in \mathbb{R}$ is the corresponding **eigenvalue**.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.
- Matrix **trace**:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

- Matrix **determinant**:

$$|A| = \det(A) = \prod_i \lambda_i$$

- Properties: $|AB| = |A||B|$, $|A^T| = |A|$, $|\alpha A| = \alpha^n |A|$

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ is **invertible** if there is $B \in \mathbb{R}^{n \times n}$ s.t. $AB = BA = I$.
- ...matrix B , such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n \times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A$, $(A^{-1})^T = (A^T)^{-1}$, $(AB)^{-1} = B^{-1}A^{-1}$
- There are many algorithms to compute A^{-1} ; general case, computational cost $O(n^3)$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n \times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^n \sum_{j=1}^n A_{i,j} x_i x_j \in \mathbb{R}$$

is called a **quadratic form**.

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive semi-definite** (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A x \geq 0$.
- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is **positive definite** (PD) if, for any $x \in \mathbb{R}^n$, ($x \neq 0$) $\Rightarrow x^T A x > 0$.
- Matrix $A \in \mathbb{R}^{n \times n}$ is PSD \Leftrightarrow all $\lambda_i(A) \geq 0$.
- Matrix $A \in \mathbb{R}^{n \times n}$ is PD \Leftrightarrow all $\lambda_i(A) > 0$.

Outline

① Introduction

② Class Administrativia

③ Recap

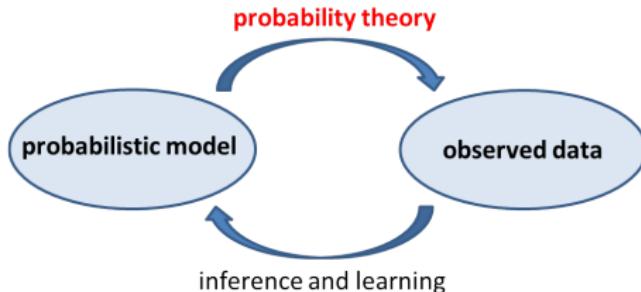
Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Probability theory



- “Essentially, all models are wrong, but some are useful”; G. Box, 1987
- The study of probability has roots in games of chance (dice, cards, ...)
- Natural tool to model uncertainty, information, knowledge, belief, ...
- ...thus also learning, inference, ...

What is probability?

- Classical definition: $\mathbb{P}(A) = \frac{N_A}{N}$

...with N mutually exclusive equally likely outcomes,
 N_A of which result in the occurrence of event A .

Laplace, 1814

Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

Example: $\mathbb{P}(\text{getting 1 in throwing a fair die}) = 1/6$.

- Frequentist definition: $\mathbb{P}(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N}$

...relative frequency of occurrence of A in infinite number of trials.

- Subjective probability: $\mathbb{P}(A)$ is a degree of belief.

de Finetti, 1930s

...gives meaning to $\mathbb{P}(\text{"Tomorrow it will rain"})$.

Key concepts: Sample space and events

- **Sample space** \mathcal{X} = set of possible outcomes of a random experiment.

Examples:

- Tossing two coins: $\mathcal{X} = \{HH, TH, HT, TT\}$
- Roulette: $\mathcal{X} = \{1, 2, \dots, 36\}$
- Draw a card from a shuffled deck: $\mathcal{X} = \{A\clubsuit, 2\clubsuit, \dots, Q\diamondsuit, K\diamondsuit\}$.
- An **event** A is a subset of \mathcal{X} : $A \subseteq \mathcal{X}$.

Examples:

- “exactly one H in 2-coin toss”: $A = \{TH, HT\} \subset \{HH, TH, HT, TT\}$.
- “odd number in the roulette”: $B = \{1, 3, \dots, 35\} \subset \{1, 2, \dots, 36\}$.
- “drawn a \heartsuit card”: $C = \{A\heartsuit, 2\heartsuit, \dots, K\heartsuit\} \subset \{A\clubsuit, \dots, K\diamondsuit\}$

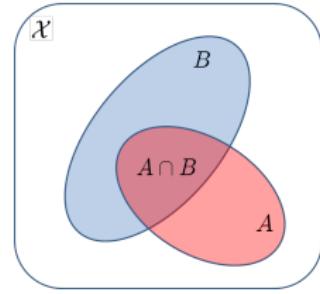
Kolmogorov's Axioms for Probability

- Probability is a function that maps events A into the interval $[0, 1]$.

Kolmogorov's axioms (1933) for probability \mathbb{P}

- For any A , $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\mathcal{X}) = 1$
- If $A_1, A_2 \dots \subseteq \mathcal{X}$ are disjoint events, then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$
- From these axioms, many results can be derived. **Examples:**

- $\mathbb{P}(\emptyset) = 0$
- $C \subset D \Rightarrow \mathbb{P}(C) \leq \mathbb{P}(D)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$



Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent ($A \perp\!\!\!\perp B$) $\Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

- Example: \mathcal{X} = “52 cards”, $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, \dots, K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52}$$

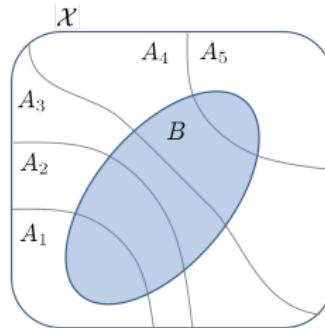
$$\mathbb{P}(A)\mathbb{P}(B) = \frac{1}{13} \cdot \frac{1}{4} = \frac{1}{52}$$

$$\mathbb{P}(A|B) = \mathbb{P}(\text{“3”} | \text{“}\heartsuit\text{”}) = \frac{1}{13} = \mathbb{P}(A)$$

Law of Total Probability and Bayes Theorem

- Law of total probability: if A_1, \dots, A_n are a partition of \mathcal{X}

$$\begin{aligned}\mathbb{P}(B) &= \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i) \\ &= \sum_i \mathbb{P}(B \cap A_i)\end{aligned}$$

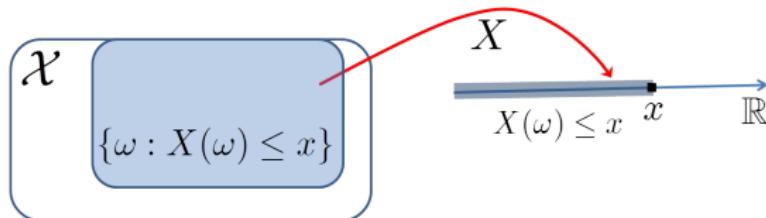


- Bayes' theorem: if $\{A_1, \dots, A_n\}$ is a partition of \mathcal{X}

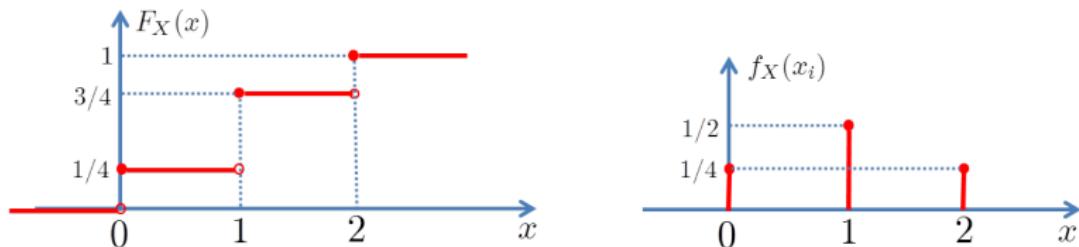
$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)}$$

Discrete Random Variables

- **Distribution function:** $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$



- **Example:** number of heads in tossing 2 coins; $\text{range}(X) = \{0, 1, 2\}$.



- **Probability mass function (discrete RV):** $f_X(x) = \mathbb{P}(X = x)$,

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i).$$

Important Discrete Random Variables

- **Uniform:** $X \in \{x_1, \dots, x_K\}$, pmf $f_X(x_i) = 1/K$.
- **Bernoulli RV:** $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \Leftarrow x = 1 \\ 1 - p & \Leftarrow x = 0 \end{cases}$

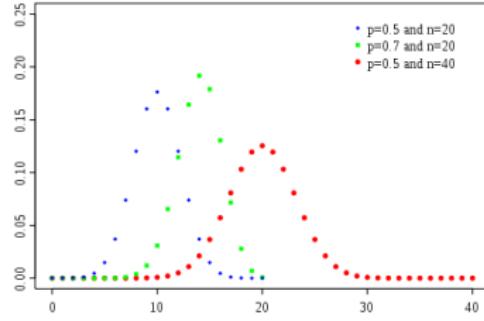
Can be written compactly as $f_X(x) = p^x(1 - p)^{1-x}$.

- **Binomial RV:** $X \in \{0, 1, \dots, n\}$ (sum on n Bernoulli RVs)

$$f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1 - p)^{(n-x)}$$

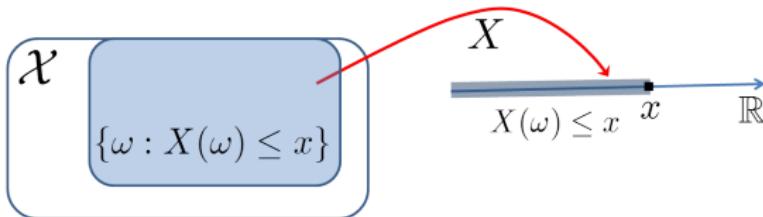
Binomial coefficients
(“ n choose x ”):

$$\binom{n}{x} = \frac{n!}{(n - x)! x!}$$

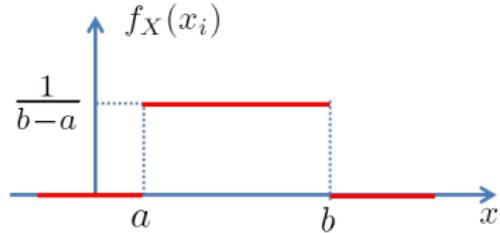
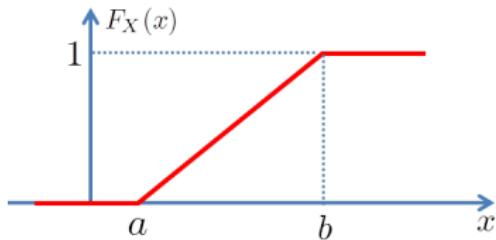


Continuous Random Variables

- Distribution function: $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$



- Example: continuous RV with uniform distribution on $[a, b]$.

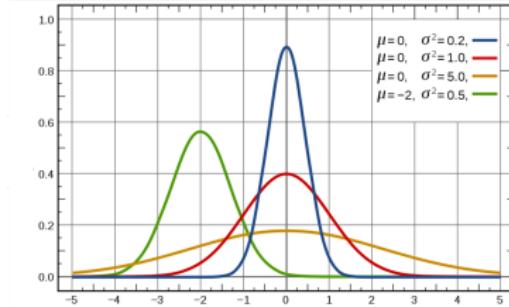


- Probability density function (pdf, continuous RV): $f_X(x)$

$$F_X(x) = \int_{-\infty}^x f_X(u) du, \quad \mathbb{P}(X \in [c, d]) = \int_c^d f_X(x) dx, \quad \mathbb{P}(X=x) = 0$$

Important Continuous Random Variables

- **Uniform:** $f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \Leftarrow x \in [a, b] \\ 0 & \Leftarrow x \notin [a, b] \end{cases}$
(previous slide).
- **Gaussian:** $f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Expectation of Random Variables

$$\bullet \text{ Expectation: } \mathbb{E}(X) = \begin{cases} \sum_i x_i f_X(x_i) & X \in \{x_1, \dots, x_K\} \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x f_X(x) dx & X \text{ continuous} \end{cases}$$

- Example: Bernoulli, $f_X(x) = p^x (1-p)^{1-x}$, for $x \in \{0, 1\}$.

$$\mathbb{E}(X) = 0(1-p) + 1p = p.$$

- Example: Gaussian, $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$. $\mathbb{E}(X) = \mu$.
- Linearity of expectation:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y); \quad \mathbb{E}(\alpha X) = \alpha \mathbb{E}(X), \quad \alpha \in \mathbb{R}$$

Expectation of Functions of Random Variables

- $\mathbb{E}(g(X)) = \begin{cases} \sum_i g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & X \text{ continuous} \end{cases}$
- **Example:** variance, $\text{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- **Example:** Bernoulli variance, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$, thus $\text{var}(X) = p(1 - p)$.
- **Example:** Gaussian variance, $\mathbb{E}((X - \mu)^2) = \sigma^2$.
- Probability as expectation of indicator, $1_A(x) = \begin{cases} 1 & \Leftarrow x \in A \\ 0 & \Leftarrow x \notin A \end{cases}$

$$\mathbb{P}(X \in A) = \int_A f_X(x) dx = \int 1_A(x) f_X(x) dx = \mathbb{E}(1_A(X))$$

Two (or More) Random Variables

- **Joint pmf** of two discrete RVs: $f_{X,Y}(x,y) = \mathbb{P}(X = x \wedge Y = y)$.

Extends trivially to more than two RVs.

- **Joint pdf** of two continuous RVs: $f_{X,Y}(x,y)$, such that

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x,y) dx dy.$$

Extends trivially to more than two RVs.

- **Marginalization:** $f_Y(y) = \begin{cases} \sum_x f_{X,Y}(x,y), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx, & \text{if } X \text{ continuous} \end{cases}$

- **Independence:**

$$X \perp\!\!\!\perp Y \Leftrightarrow f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad \begin{matrix} \Rightarrow \\ \not= \end{matrix} \quad \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Conditionals and Bayes' Theorem

- **Conditional pmf** (discrete RVs):

$$f_{X|Y}(x|y) = \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x \wedge Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

- **Conditional pdf** (continuous RVs): $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
- **Bayes' theorem**: $f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}$ (pdf or pmf).
- Also valid in the mixed case (e.g., X continuous, Y discrete).

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$, $f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$,

$$f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}, \quad f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}.$$

- Conditional probabilities:

$f_{X Y}(x y)$	$Y = 0$	$Y = 1$
$X = 0$	2/3	4/7
$X = 1$	1/3	3/7

$f_{Y X}(y x)$	$Y = 0$	$Y = 1$
$X = 0$	1/3	2/3
$X = 1$	1/4	3/4

An Important Multivariate RV: Multinomial

- **Multinomial:** $X = (X_1, \dots, X_K)$, $X_i \in \{0, \dots, n\}$, such that $\sum_i X_i = n$,

$$f_X(x_1, \dots, x_K) = \begin{cases} \binom{n}{x_1 \ x_2 \ \dots \ x_K} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \Leftrightarrow \sum_i x_i = n \\ 0 & \Leftrightarrow \sum_i x_i \neq n \end{cases}$$

$$\binom{n}{x_1 \ x_2 \ \dots \ x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

Parameters: $p_1, \dots, p_K \geq 0$, such that $\sum_i p_i = 1$.

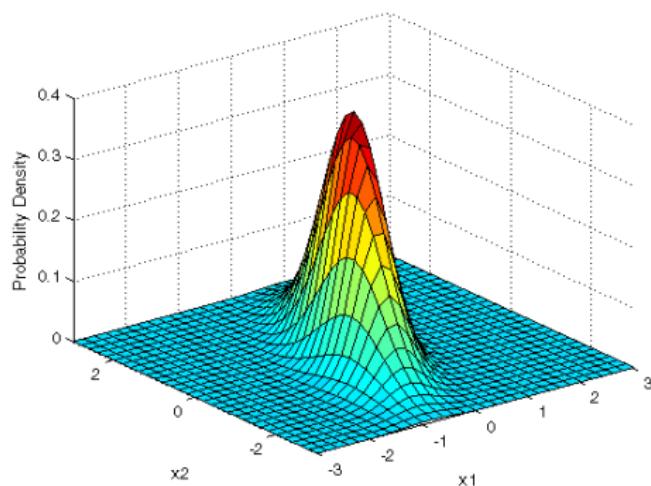
- Generalizes the binomial from binary to K -classes.
- Example: tossing n independent fair dice, $p_1 = \cdots = p_6 = 1/6$.
 x_i = number of outcomes with i dots. Of course, $\sum_i x_i = n$.
- For $n = 1$, sometimes called **categorical** or **multinoulli**.
For $n = 1$, one and only one $x_i = 1$, others are 0, thus $\binom{n}{x_1 \ \dots \ x_K} = 1$.

An Important Multivariate RV: Gaussian

- Multivariate Gaussian: $X \in \mathbb{R}^n$,

$$f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left(-\frac{1}{2}(x - \mu)^T C^{-1}(x - \mu)\right)$$

- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n \times n}$.
Expected value: $\mathbb{E}(X) = \mu$. Meaning of C : next slide.



Covariance, Correlation, and all that...

- **Covariance** between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

- Relationship with variance: $\text{var}(X) = \text{cov}(X, X)$.
- $X \perp\!\!\!\perp Y \Leftrightarrow f_{X,Y}(x,y) = f_X(x)f_Y(y) \stackrel{\Rightarrow}{\cancel{\Leftarrow}} \text{cov}(X, Y) = 0$
- **Covariance matrix** of multivariate RV, $X \in \mathbb{R}^n$:
$$\text{cov}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\right] = \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T$$
- Covariance of Gaussian RV, $f_X(x) = \mathcal{N}(x; \mu, C) \Rightarrow \text{cov}(X) = C$

More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = ACA^T$;
- If $\text{cov}(X) = C$ and $Y = a^T X \in \mathbb{R}$, then $\text{var}(Y) = a^T Ca \geq 0$;
- If $\text{cov}(X) = C$ and $Y = C^{-1/2}X$, then $\text{cov}(Y) = I$;
- If $f_X(x) = \mathcal{N}(x; 0, I)$ and $Y = \mu + C^{1/2}X$, then $f_Y(y) = \mathcal{N}(y; \mu, C)$;
- If $f_X(x) = \mathcal{N}(x; \mu, C)$ and $Y = C^{-1/2}(X - \mu)$, then $f_Y(y) = \mathcal{N}(y; 0, I)$.

Entropy and all that...

Entropy of a discrete RV $X \in \{1, \dots, K\}$:

$$H(X) = - \sum_{x=1}^K f_X(x) \log f_X(x)$$

- **Positivity**: $H(X) \geq 0$;
 $H(X) = 0 \Leftrightarrow f_X(i) = 1$, for exactly one $i \in \{1, \dots, K\}$.
- **Upper bound**: $H(X) \leq \log K$;
 $H(X) = \log K \Leftrightarrow f_X(x) = 1/k$, for all $x \in \{1, \dots, K\}$
- Measure of **uncertainty/randomness** of X

Continuous RV X , differential entropy:

$$h(X) = - \int f_X(x) \log f_X(x) dx$$

- $h(X)$ can be positive or negative. Example, if $f_X(x) = \text{Uniform}(x; a, b)$, $h(X) = \log(b - a)$.
- If $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$, then $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$.

Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

$$D(f_X \| g_X) = \sum_{x=1}^K f_X(x) \log \frac{f_X(x)}{g_X(x)}$$

Positivity: $D(f_X \| g_X) \geq 0$

$$D(f_X \| g_X) = 0 \Leftrightarrow f_X(x) = g_X(x), \text{ for } x \in \{1, \dots, K\}$$

KLD between two pdf:

$$D(f_X \| g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} dx$$

Positivity: $D(f_X \| g_X) \geq 0$

$$D(f_X \| g_X) = 0 \Leftrightarrow f_X(x) = g_X(x), \text{ almost everywhere}$$

Recommended Reading

- A. Maleki and T. Do, “Review of Probability Theory”, Stanford University, 2017 (<https://tinyurl.com/pz7p9g5>)
- L. Wasserman, “All of Statistics: A Concise Course in Statistical Inference”, Springer, 2004.

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Minimizing a function

- We are given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.
- Goal: find x^* that minimizes $f : \mathbb{R}^n \rightarrow \mathbb{R}$.
- **Global minimum**: for any $x \in \mathbb{R}^n$, $f(x^*) \leq f(x)$.
- **Local minimum**: for any $\|x - x^*\| \leq \delta \Rightarrow f(x^*) \leq f(x)$.
- Are local minima also global minima?

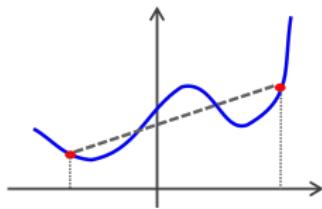
Convex Functions

- Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$
- f is a **convex function**, if, for any $\lambda \in [0, 1]$, and any x, x' ,

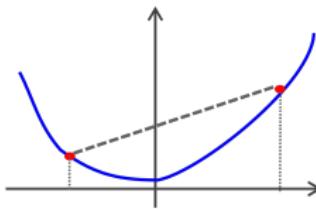
$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x')$$

- f is a **strictly convex function**, if, for any $\lambda \in]0, 1[$, and any x, x' ,

$$f(\lambda x + (1 - \lambda)x') < \lambda f(x) + (1 - \lambda)f(x')$$

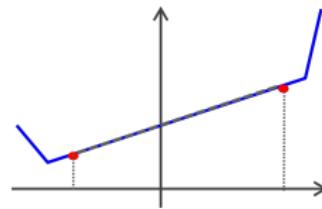


non-convex



convex

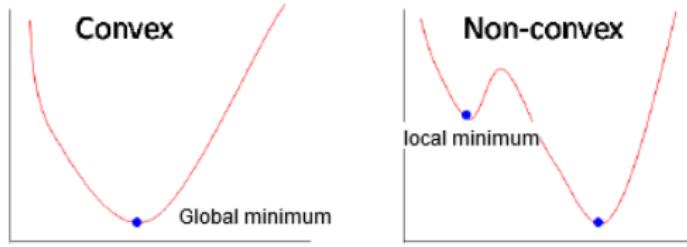
strictly convex



convex, not strictly

Relationship Between Convexity and Minimization

- Goal: find x^* that minimizes $f : \mathbb{R}^n \rightarrow \mathbb{R}$.
- If f is convex and x^* is a local minimizer, then it is also a global minimizer.
- If f is strictly convex and x^* is a local minimizer, then it is also the unique global minimizer.



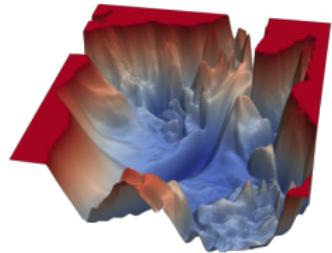
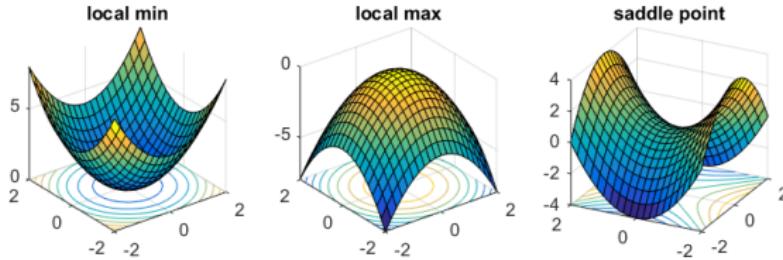
Gradients and Minimization

- Given $f : \mathbb{R}^n \rightarrow \mathbb{R}$ (differentiable), the **gradient** of f at x :

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

- Relationship between gradient and minimization

$$x^* \text{ is local minimizer} \stackrel{\Rightarrow}{\cancel{\Rightarrow}} \nabla f(x^*) = 0$$



Hessians and Convexity

- Given $f : \mathbb{R}^n \rightarrow \mathbb{R}$ (differentiable), the **Hessian** of f at x :

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

- Relationship between Hessian and convexity:

- ✓ Positive semi-definite Hessian $\Leftrightarrow f$ is **convex**
- ✓ Positive definite Hessian $\Leftrightarrow f$ is **strictly convex**.

More on Gradients

- Gradient of quadratic form $f(x) = x^T Ax$: $\nabla f(x) = (A + A^T)x$
- ...if A symmetric: $\nabla f(x) = 2Ax$
- .Particular case: $f(x) = x^T x = \|x\|_2^2$, then $\nabla f(x) = 2x$
- If $f(x) = x^T b = b^T x$, then $\nabla f(x) = b$
- If $g(x) = f(Ax)$, then $\nabla g(x) = A^T \nabla f(Ax)$
- If $g(x) = f(a \odot x)$, then $\nabla g(x) = a \odot \nabla f(a \odot x)$
- In simple cases, we can find minima analytically: $f(x) = \|Ax - y\|_2^2$

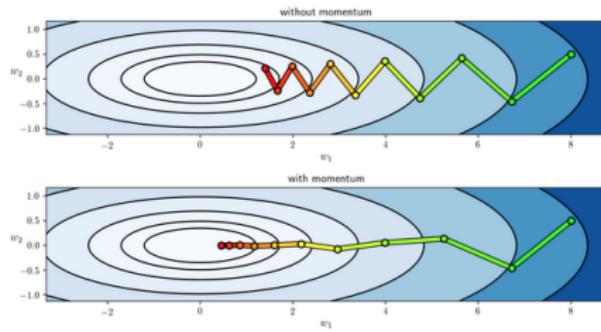
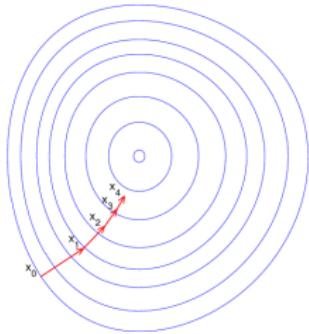
$$\nabla f(x) = 2A^T(Ax - y) = 0 \Rightarrow x^* = (A^T A)^{-1} A^T y$$

Gradient Descent

- Goal: minimize $f : \mathbb{R}^d \rightarrow \mathbb{R}$, for differentiable f
- Take **small steps** in the **negative gradient direction** until a **stopping criterion** is met:

$$x^{(t+1)} \leftarrow x^{(t)} - \eta(t) \nabla f(x^{(t)})$$

- Choosing the **step-size**: crucial for convergence and performance.
- GD may work well, or not so well. There are many ways to improve it.



Recommended Reading

- Z. Kolter and C. Do, “Linear Algebra Review and Reference”, Stanford University, 2015 (<https://tinyurl.com/44x2qj4>)

Outline

① Introduction

② Class Administrativia

③ Recap

Linear Algebra

Probability Theory Refresher

Optimization

④ Introduction to Machine Learning

Machine Learning

Tom Mitchell's definition:

- “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 .”
- In a nutshell: **learn from data; improve performance with experience**

This comes in many flavours:

- Supervised learning
- Unsupervised learning
- Self-supervised learning
- Reinforcement learning
- Active learning

Formulate the problem; get data; learn the model from the data; evaluate.

Example Tasks

Binary classification: given an e-mail: is it spam or not-spam?

Multi-class classification: given a news article, determine its topic (politics, sports, etc.)

Regression: how much time a person will spend reading this article?

 **AlphaGo Beats Go Human Champ:
Godfather Of Deep Learning Tells Us Do
Not Be Afraid Of AI**

21 March 2016, 10:16 am EDT By Aaron Mamiit Tech Times



Last week, Google's artificial intelligence program

Last week, Google's artificial intelligence program AlphaGo **dominated** its match with South Korean world Go champion Lee Sedol, winning with a 4-1 score.

The achievement stunned artificial intelligence experts, who previously thought that Google's computer program would need at least 10 more years before developing enough to be able to beat a human world champion.



sports
politics
technology
economy
weather
culture

Let's Start Simple

- Example 1 – sequence: $\star \diamond \circ$; label: -1
- Example 2 – sequence: $\star \heartsuit \triangle$; label: -1
- Example 3 – sequence: $\star \triangle \spadesuit$; label: $+1$
- Example 4 – sequence: $\diamond \triangle \circ$; label: $+1$
- New sequence: $\star \diamond \circ$; label -1
- New sequence: $\star \diamond \heartsuit$; label -1
- New sequence: $\star \triangle \circ$; label ?

Why can we do this?

(Credits: Ryan McDonald)

Let's Start Simple: Machine Learning

- Example 1 – sequence: $\star \diamond \circ$; label: -1
- Example 2 – sequence: $\star \heartsuit \triangle$; label: -1
- Example 3 – sequence: $\star \triangle \spadesuit$; label: $+1$
- Example 4 – sequence: $\diamond \triangle \circ$; label: $+1$
- New sequence: $\star \diamond \heartsuit$; label -1

Label -1

$$P(-1|\star) = \frac{\text{count}(\star \text{ and } -1)}{\text{count}(\star)} = \frac{2}{3} = 0.67 \text{ vs. } P(+1|\star) = \frac{\text{count}(\star \text{ and } +1)}{\text{count}(\star)} = \frac{1}{3} = 0.33$$

$$P(-1|\diamond) = \frac{\text{count}(\diamond \text{ and } -1)}{\text{count}(\diamond)} = \frac{1}{2} = 0.5 \text{ vs. } P(+1|\diamond) = \frac{\text{count}(\diamond \text{ and } +1)}{\text{count}(\diamond)} = \frac{1}{2} = 0.5$$

$$P(-1|\heartsuit) = \frac{\text{count}(\heartsuit \text{ and } -1)}{\text{count}(\heartsuit)} = \frac{1}{1} = 1.0 \text{ vs. } P(+1|\heartsuit) = \frac{\text{count}(\heartsuit \text{ and } +1)}{\text{count}(\heartsuit)} = \frac{0}{1} = 0.0$$

Label $+1$

(Credits: Ryan McDonald)

Let's Start Simple: Machine Learning

- Example 1 – sequence: $\star \diamond \circ$; label: -1
- Example 2 – sequence: $\star \heartsuit \triangle$; label: -1
- Example 3 – sequence: $\star \triangle \spadesuit$; label: $+1$
- Example 4 – sequence: $\diamond \triangle \circ$; label: $+1$
- New sequence: $\star \triangle \circ$; label ?

Label -1

Label $+1$

$$P(-1|\star) = \frac{\text{count}(\star \text{ and } -1)}{\text{count}(\star)} = \frac{2}{3} = 0.67 \text{ vs. } P(+1|\star) = \frac{\text{count}(\star \text{ and } +1)}{\text{count}(\star)} = \frac{1}{3} = 0.33$$

$$P(-1|\triangle) = \frac{\text{count}(\triangle \text{ and } -1)}{\text{count}(\triangle)} = \frac{1}{3} = 0.33 \text{ vs. } P(+1|\triangle) = \frac{\text{count}(\triangle \text{ and } +1)}{\text{count}(\triangle)} = \frac{2}{3} = 0.67$$

$$P(-1|\circ) = \frac{\text{count}(\circ \text{ and } -1)}{\text{count}(\circ)} = \frac{1}{2} = 0.5 \text{ vs. } P(+1|\circ) = \frac{\text{count}(\circ \text{ and } +1)}{\text{count}(\circ)} = \frac{1}{2} = 0.5$$

(Credits: Ryan McDonald)

Machine Learning

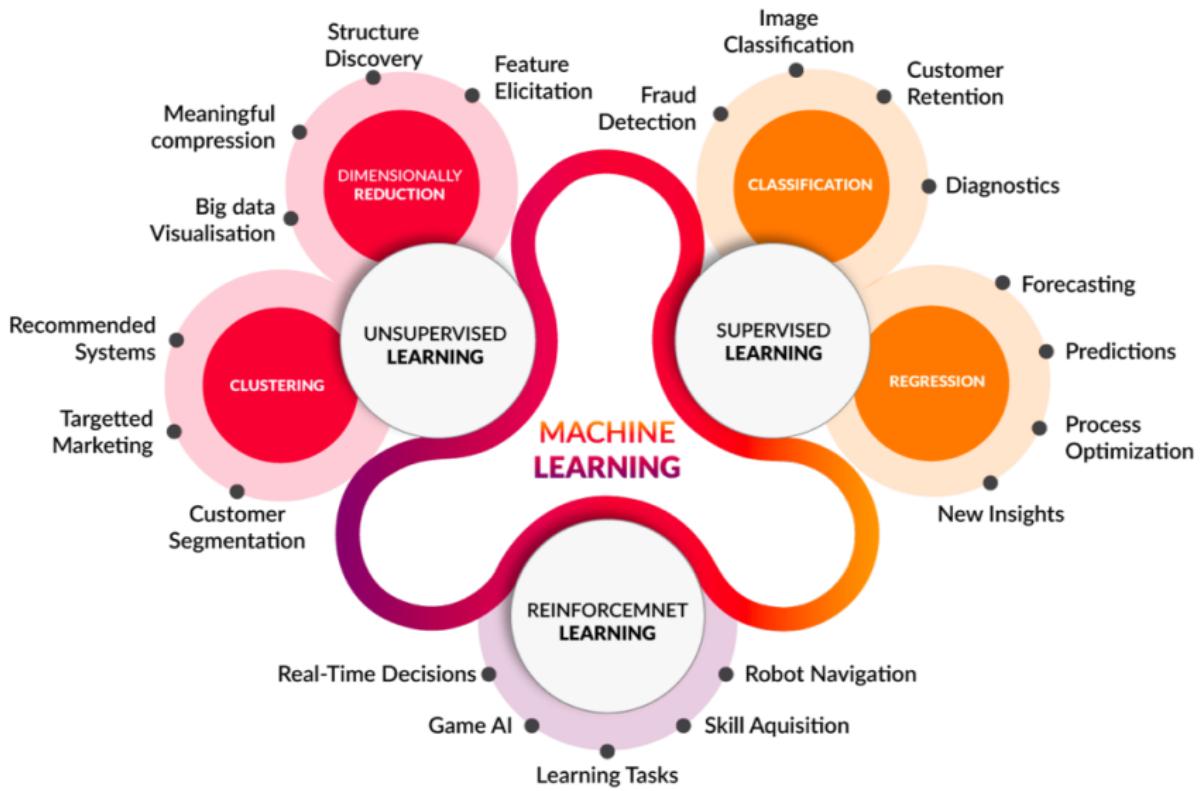
- ① Define a model/distribution of interest
 - ② Make some assumptions if needed
 - ③ Fit the model to the data
- Model: $P(\text{label}|\text{sequence}) = P(\text{label}|\text{symbol}_1, \dots, \text{symbol}_n)$
 - Prediction for new sequence = $\text{argmax}_{\text{label}} P(\text{label}|\text{sequence})$
 - Assumption (**naive Bayes**):
$$P(\text{symbol}_1, \dots, \text{symbol}_n | \text{label}) = \prod_{i=1}^n P(\text{symbol}_i | \text{label})$$
 - Fit the model to the data: count!! (simple probabilistic modeling)

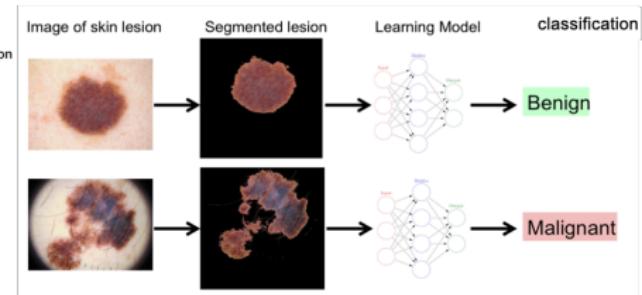
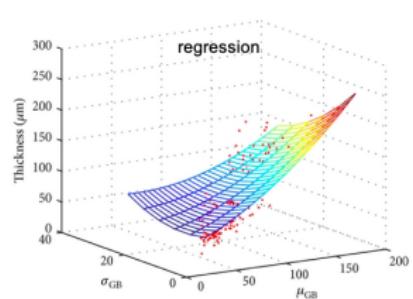
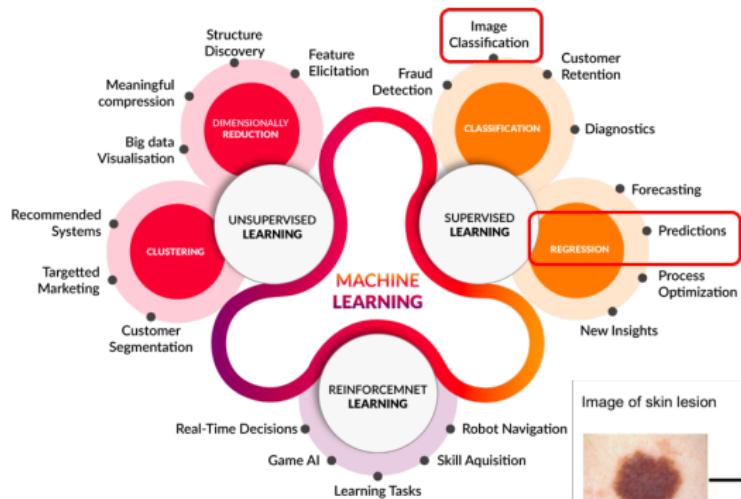
Some Notation: Inputs and Outputs

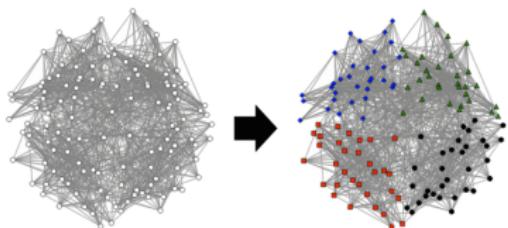
- Input $x \in \mathcal{X}$
 - e.g., a news article, a sentence, an image, ...
- Output $y \in \mathcal{Y}$
 - e.g., spam/not spam, a topic, the object in the image (cat? dog?); a segmentation of the image (pedestrian; car; grass; background)
- Input/Output pair: $(x, y) \in \mathcal{X} \times \mathcal{Y}$
 - e.g., a **news article** together with a **topic**
 - e.g., a **image** together with an **object**
 - e.g., an **image** partitioned into **segmentation regions**

Many Flavours

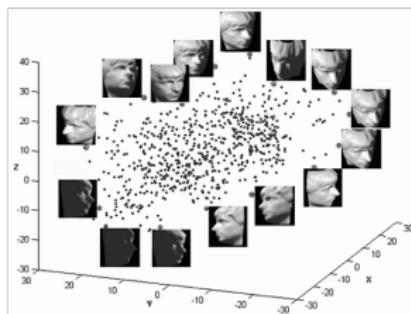
- **Supervised learning:** pairs (x, y) are provided at training time (the main focus of this class)
 - Examples: perceptron, SVMs, decision trees, nearest neighbor, neural networks, ...
 - Caveat: the labels y may be hard or expensive to annotate
- **Unsupervised learning:** only x is provided; the model needs to figure out what the labels are without any supervision
 - Examples: clustering, PCA, ...
- **Reinforcement learning:** x is provided, and the model can act on the environment to obtain a reward (but doesn't get to know y)
 - Example: a robot acting on an environment trying to achieve a goal
- **Active learning:** the model requests which data points to label next.



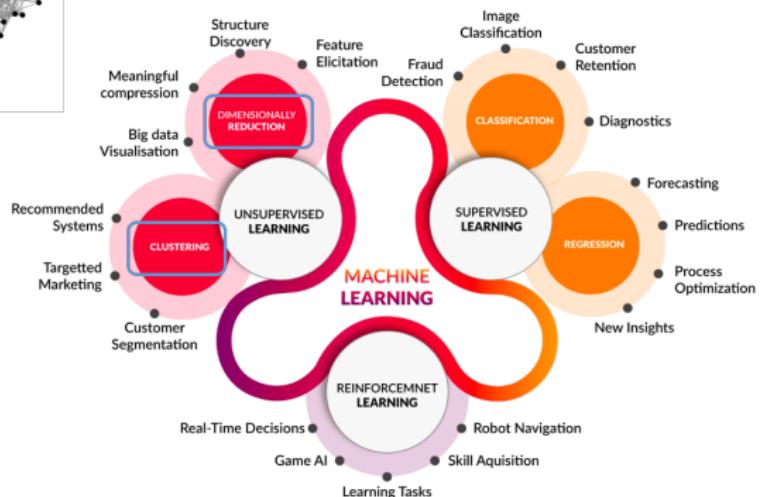


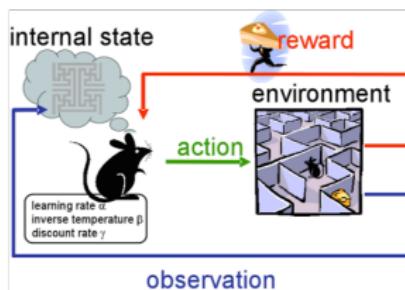
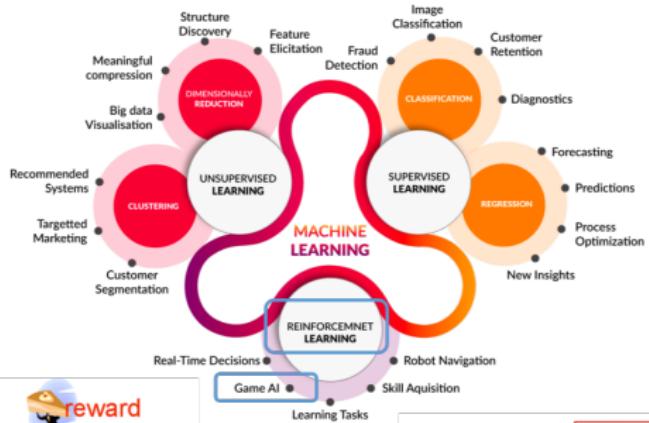


Community extraction from networks

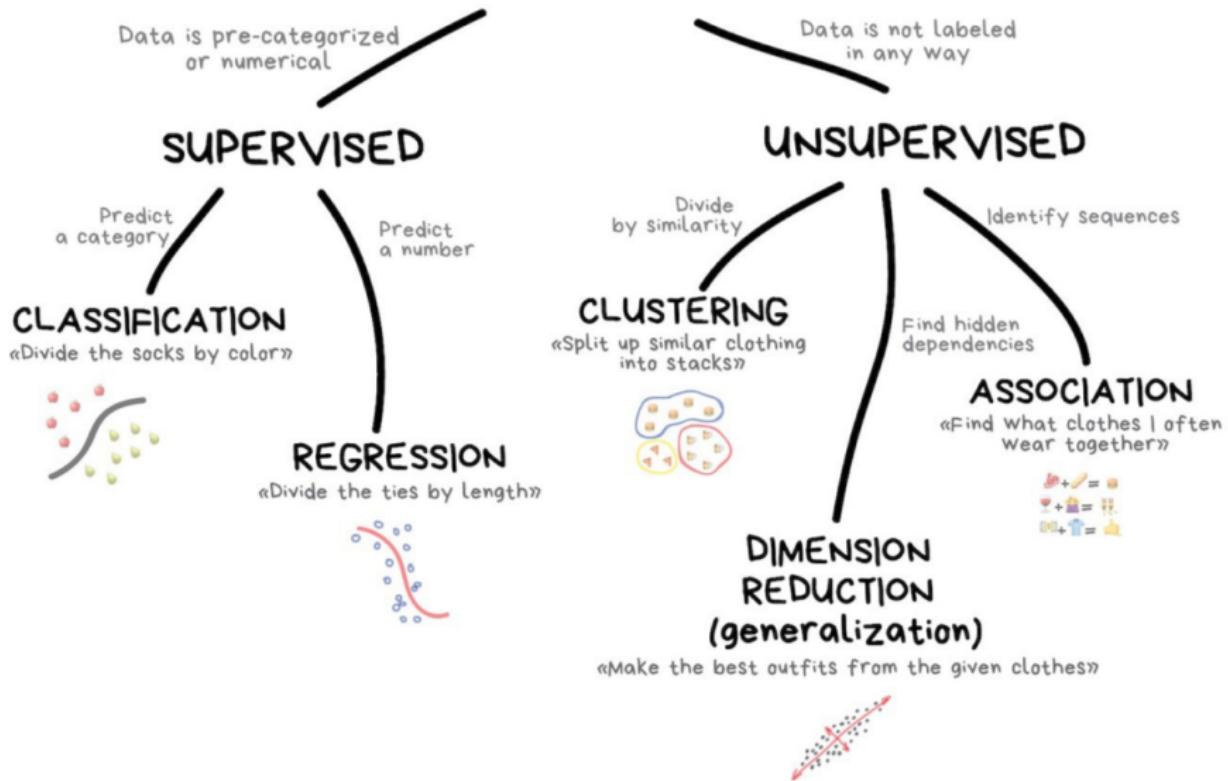


Manifold learning





CLASSICAL MACHINE LEARNING



Supervised Learning

- We are given a **labeled dataset** of input/output pairs:

$$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$$

- **Goal:** learn a **predictor** $h : \mathcal{X} \rightarrow \mathcal{Y}$ that generalizes well to arbitrary inputs.
- At test time, given $x \in \mathcal{X}$, we predict

$$\hat{y} = h(x).$$

- Hopefully, $\hat{y} \approx y$ most of the time.

Tasks/problems have different names depending on what \mathcal{Y} is..

Regression deals with **continuous** output variables:

- Simple **regression**: $\mathcal{Y} = \mathbb{R}$
 - e.g., given a news article, how much time a user will spend reading it?
- **Multivariate regression**: $\mathcal{Y} = \mathbb{R}^K$
 - e.g., predict the X-Y coordinates in an image where the user will click

Classification

Tasks/problems have different names depending on what \mathcal{Y} is...

Classification deals with **discrete** output variables:

- **Binary classification:** $\mathcal{Y} = \{\pm 1\}$
 - e.g., spam detection
- **Multi-class classification:** $\mathcal{Y} = \{1, 2, \dots, K\}$
 - e.g., topic classification
- **Structured classification:** \mathcal{Y} exponentially large and structured
 - e.g., machine translation, caption generation, image segmentation
 - **Later in this course!**

Sometimes **reductions** are convenient:

- logistic regression reduces classification to regression
- one-vs-all reduces multi-class to binary
- greedy search reduces structured classification to multi-class

... but other times it's better to tackle the problem in its native form.

More later!

Conclusions

- Machine learning allows computer programs to learn from data (observations, interventions, ...) and improve performance with experience
- Can be supervised, unsupervised, self-supervised, reinforced, etc.
- Tasks can be (binary or multi-class) classification, regression, or more nuanced

Thank you!

Questions?

