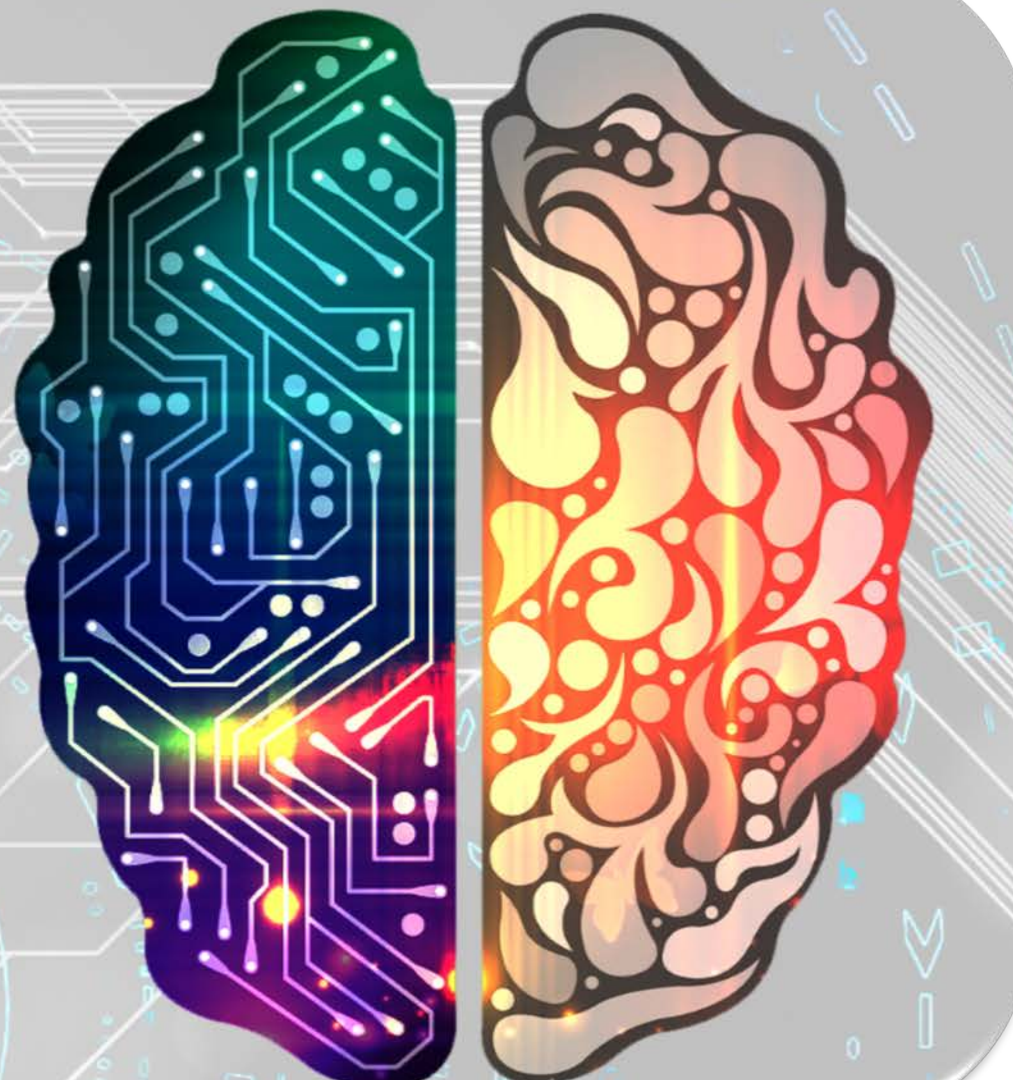
جامعة قطر
QATAR UNIVERSITY

CMPS 460 – Spring 2022

# MACHINE LEARNING

Tamer Elsayed





8.a

## Probabilistic Modeling: Review of Statistical Principles

Sec 9.1-9.3

# Probabilistic Modelling

***Learning***
**as a problem of *statistical inference***

- Describe the **training data** as a probability distribution $D$.
- **Learning** is then to infer the "best" values of the parameters $\theta$ of $D$ given the observed training data.

# Review of Some Statistical Principles

# Bayes Rule

*joint distribution*

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad \text{Bayes' rule}$$

**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

# Bayes Rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

# Joint Distribution

Showing a probability distribution for two (or more) random variables

| gender | hours_worked | wealth | |
|---|---|---|---|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

# Using a Joint Distribution

| gender | hours_worked | wealth | |
|---|---|---|---|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Given the joint distribution, we can find the probability of any logical expression E involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

# Using a Joint Distribution

| gender | hours_worked | wealth | |
|---|---|---|---|
| Female | v0:40.5- | poor | 0.253122 |
| | | rich | 0.0245895 |
| | v1:40.5+ | poor | 0.0421768 |
| | | rich | 0.0116293 |
| Male | v0:40.5- | poor | 0.331313 |
| | | rich | 0.0971295 |
| | v1:40.5+ | poor | 0.134106 |
| | | rich | 0.105933 |

Given the joint distribution, we can make ***inferences***

- e.g., P(Male | Poor)?
- or P(Wealth | Gender, Hours)?

# Recall: Classification

**If we have access to $D$ (the data generating *joint* distribution), finding an optimal classifier would be trivial!**

**We don't have access to $D$!**
**So let's try to estimate it instead!**

# "Training" in Probabilistic Settings …

**Training =**

**estimating $\mathcal{D}$ from a finite training set**

- We typically assume that $\mathcal{D}$ comes from a specific family of probability distributions
  - e.g., Bernoulli, Gaussian, etc.

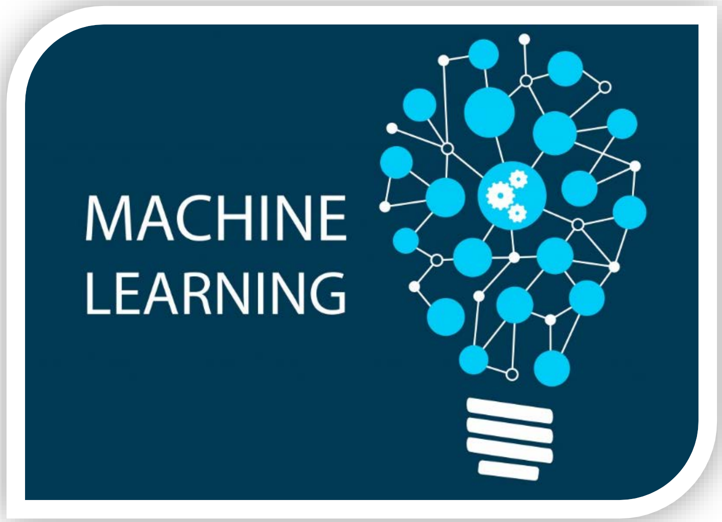
- Learning means inferring parameters of that distribution
  - e.g., mean and covariance of the Gaussian.

# Assumption: Training Examples are iid

- Independently and Identically-distributed
  - i.e., as we draw a sequence of examples from $\mathcal{D}$, the n$^{th}$ draw is independent from the previous n-1 samples.

- This assumption is usually false!
  - But sufficiently close to true to be useful

**How can we estimate the joint probability distribution from data?**




# Maximum Likelihood Estimation

# Maximum Likelihood Estimation

***Find the parameters that maximize the probability of the data***

# Experiment: a "Biased" Coin!


- Given a "biased" coin
- Flipped multiple times
- Got $\alpha_1$ heads and $\alpha_0$ tails

***What is p(head) and p(tail)?***

# Model one flip: Bernoulli


X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

Each coin flip yields a Boolean value for X

X ~ Bernoulli: $P(X) = \theta^{X}(1-\theta)^{(1-X)}$

# Multiple flips


X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

X ~ Bernoulli: $P(X) = \theta^{X}(1-\theta)^{(1-X)}$

Given a data set D of iid flips: $\alpha_1$ 1s and $\alpha_0$ 0s:

$$P_{\theta}(D) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

# Maximum Likelihood Estimation

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

$$P_\theta(D) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}_{MLE} = argmax_\theta P_\theta(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

# Maximum Likelihood Estimation


- Example: how to model a k-sided die?

  X ~ *Discrete*: $P(X) = \prod_k \theta_k{}^{1(x=k)}$

- Given a data set D of iid rolls, where

$$P_\theta(D) = \prod_{i=1}^{k} \theta_i{}^{x_i}$$

$$\hat{\theta}_{i,MLE} = \frac{x_i}{\sum_{i=1}^{K} x_i}$$