Data Mining & Machine Learning

CS37300

Profs Tianyi Zhang and Rajiv Khanna

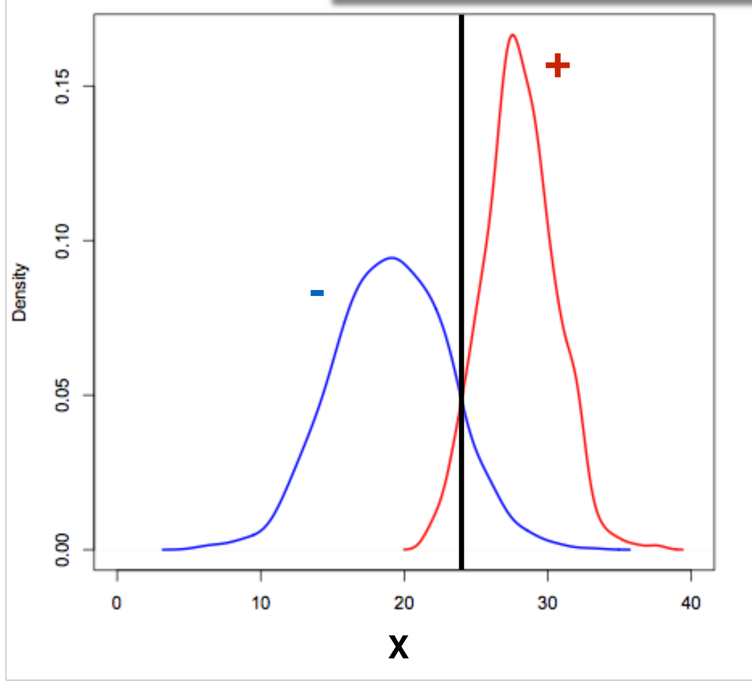
Aug 28, 2023

From last time: Example learning problem

Task: Devise a rule to classify items based on the attribute **X**

Knowledge representation: If-then rules

Example rule:
If x > 25 then +
Else -



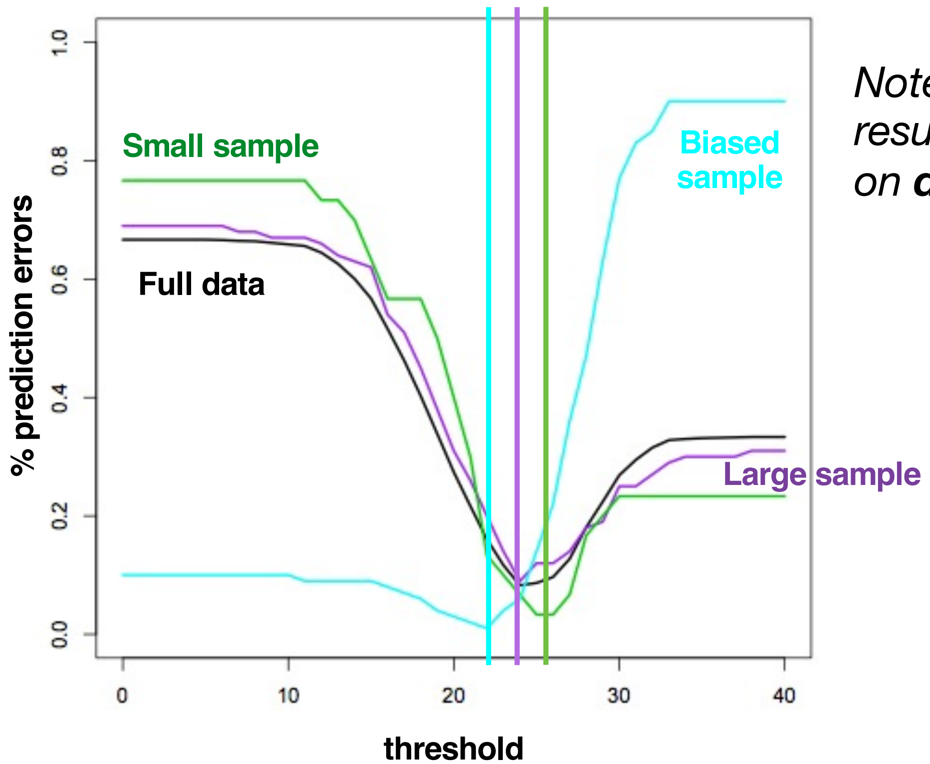
What is the model space?

All possible thresholds

What score function?

Prediction error rate

Sampling errors



Note: learning result depends on ***data***

Probability refresher

- Probability basics
  - Random variable
  - Joints and Marginals
  - Conditionals
- Independence
- Expectation, variance
- Common distributions
- Maximum likelihood estimation

Probability basics

- Random variable (RV)
  - A function/mapping that maps a set of possible outcomes (of an experiment) to a space that “can be assigned a measure of likelihood”
  - Notation: X may be a RV, x is an instance of the RV
- Types
  - Discrete (including Boolean)
  - Continuous

Basics

- Sample space
  - Domain of the RV
  - Set of all possible outcomes of an experiment that defines the RV
- Event
  - A subset of the sample space
  - Mutually exclusive events: can not occur together, i.e. intersection of the both the corresponding subsets is null

Examples

- Experiment
- One coin toss
- Two coin tosses
- Draw a card from a deck
- Play a chess game
- Rain, IsRoadWet

- Sample space
- {H,T}
- {HH, HT,TT,TH}
- 52 choices
- {Win, Lose, Draw}
- {TT,FF,TF,FT}

Two coin toss

- Sample space {HH, HT,TH,TT}
- Event A : Atleast one head {HH, HT,TH}
- Event B : Heads in first toss {HH, HT}
- Event C : Both tails
- Are A, B mutually exclusive? What about B,C ?

What is (mathematically) a probability distribution?

- For any event A from the sample space
  - $0 \leq P(A) \leq 1$
  - $\sum_A P(A) = 1$
  - $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) +$ ... if A_i are all mutually exclusive

Can be derived from axioms

- $P(A^c) = 1 - P(A)$
- $P(true) = 1, \;\; P(False) = 0$
- If A, B are mutually exclusive, $P(A \cap B) = ??$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Calculating probability

- Probability = $\frac{\#\ favorable\ outcomes}{\#possible\ outcomes}$
- Example:
  - Roll a fair 6 sided dice twice, what is the probability that sum of the two rolls is 8?
  - Total #possible outcomes = 6 *6 = 36
  - Favorable outcomes: {2,6},{3,5},{4,4},{5,3},{6,2}
  - Probability 5/36.

Probability distribution

- Probability distribution: The function specifying the probability of every realization of the random variable

- Discrete: P(X=x)

- Continuous: Probability of any singular event is 0, but
  - $P(a < X < b) = \int_a^b p(x)dx$

Joint distribution

- Joint distribution specifies the probability of every joint realization of two or more random variables.
- Example: Two coin tosses, two dice throws

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Marginal

- Marginal probability: Probability of an event for a random variable irrespective of realizations of other random variables

- $P(A) = \sum_b P(A, B = b)$

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

P(weather=cloudy) = 0.31 + 0.02

Conditional probability

- Probability of an event GIVEN another event has already happened
- E.g. P(A | A) = ?
- E.g. P(A | B) = ? If A, B are mutually exclusive?
- Generally: $P(A \mid B) = \frac{P(A,B)}{P(B)}$

Conditional (contd.)

- What is P(warning = Y | weather = snow) ?

	weather = sunny	weather = rainy	weather = cloudy	weather= snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- P (warning = Y, weather = snow) = 0.02
- P(weather = snow) = 0.02 + 0.03 = 0.05
- P(warning = Y | weather = snow) = 0.02/0.05

Independence

- Events A and B are independent iff
  - $P(A \wedge B) = P(A)\, P(B)$
  - Equivalently: $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$
  - *Knowing B happens tells you nothing about whether A happens*
- Random variables X and Y are independent iff
  every event about X is independent of every event about Y.
  - Equivalently: joint distribution $P_{(X,Y)}$ is equal $P_X P_Y$ product of marginal distributions
  - If discrete variables: $P(Y=y,X=x) = P(Y=y)P(X=x)$, or $P(Y=y|X=x) = P(Y=y)$
- Examples
  - Coin flip 1 and coin flip 2?
  - Weather and storm warning?
  - Weather and coin flip=H?

Example of independent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.15	0.075	→ P(X=1) = 0.25
X = 2	0.075	0.45	0.225	→ P(X=2) = 0.75
	↓ P(Y=1) = 0.1	↓ P(Y=2) = 0.6	↓ P(Y=3) = 0.3	

- P(X=1,Y=1) = P(X=1) P(Y=1) ? P(X=2,Y=1) = P(X=2) P(Y=1) ?
- P(X=1,Y=2) = P(X=1) P(Y=2) ? P(X=2,Y=2) = P(X=2) P(Y=2) ?
- P(X=1,Y=3) = P(X=1) P(Y=3) ? P(X=2,Y=3) = P(X=2) P(Y=3) ?
- If the answer to the 6 questions above is “Yes”, then X and Y are independent

Example of independent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.15	0.075	P(X=1) = 0.25
X = 2	0.075	0.45	0.225	P(X=2) = 0.75
	P(Y=1) = 0.1	P(Y=2) = 0.6	P(Y=3) = 0.3	

- 0.025 = 0.25 * 0.1 (Yes)
- 0.15 = 0.25 * 0.6 (Yes)
- 0.075 = 0.25 * 0.3 (Yes)

0.075 = 0.75 * 0.1 (Yes)

0.45 = 0.75 * 0.6 (Yes)

0.225 = 0.75 * 0.3 (Yes)

- The answer to the 6 questions above is “Yes”. **X and Y are independent.**

Example of independent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3
X = 1	0.025	0.15	0.075
X = 2	0.075	0.45	0.225

P(X=1) = 0.25

P(X=2) = 0.75

P(Y=1) = 0.1 P(Y=2) = 0.6 P(Y=3) = 0.3

- Quick way to check:
  - In every column, values have proportions 1:3
  - So conditional distribution X given Y=y doesn't depend on y
  - P(X=x|Y=y) = P(X=x)

Example of dependent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.125	0.1	P(X=1) = 0.25
X = 2	0.075	0.475	0.2	P(X=2) = 0.75
	P(Y=1) = 0.1	P(Y=2) = 0.6	P(Y=3) = 0.3	

- P(X=1,Y=1) = P(X=1) P(Y=1) ? P(X=2,Y=1) = P(X=2) P(Y=1) ?
- P(X=1,Y=2) = P(X=1) P(Y=2) ? P(X=2,Y=2) = P(X=2) P(Y=2) ?
- P(X=1,Y=3) = P(X=1) P(Y=3) ? P(X=2,Y=3) = P(X=2) P(Y=3) ?
- If the answer to the 6 questions above is “Yes”, then X and Y are independent

Example of dependent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.125	0.1	→ P(X=1) = 0.25
X = 2	0.075	0.475	0.2	→ P(X=2) = 0.75
	↓ P(Y=1) = 0.1	↓ P(Y=2) = 0.6	↓ P(Y=3) = 0.3	

- 0.025 = 0.25 * 0.1 (Yes) 0.075 = 0.75 * 0.1 (Yes)
- 0.125 = 0.25 * 0.6 (No) 0.475 = 0.75 * 0.6 (No)
- 0.1 = 0.25 * 0.3 (No) 0.2 = 0.75 * 0.3 (No)

- The answer to at least 1 question above is “No”. **X and Y are NOT independent.**

Example of dependent variables

- How to check for independence?
- Joint probability P(X,Y)

	Y = 1	Y = 2	Y = 3	
X = 1	0.025	0.125	0.1	→ P(X=1) = 0.25
X = 2	0.075	0.475	0.2	→ P(X=2) = 0.75
	↓ P(Y=1) = 0.1	↓ P(Y=2) = 0.6	↓ P(Y=3) = 0.3	

- First column has proportions 1:3
- Third column has proportions 1:2
- P(X=x|Y=y) depends on y.
- They can’t be independent.

Mutual Independence

- Multiple events $A_1, A_2, \ldots A_n$ are **(mutually) independent** iff
- Every $I \subset \{1,2,\ldots,n\}$ and $J \subset \{1,2,\ldots,n\}$ have

$$P\left(\bigcap_{i\in I} A_i \cap \bigcap_{j\in J} A_j\right) = P\left(\bigcap_{i\in I} A_i\right) P\left(\bigcap_{j\in J} A_j\right)$$

- Random variables $X_1, X_2, \ldots X_n$ are (mutually) independent iff
- Every event A_1 about X_1, event A_2 about $X_2, \ldots$ event A_n about X_n
- satisfy that $A_1, A_2, \ldots, A_n$ are mutually independent

Conditional independence

- Two events *A* and *B* are **conditionally** independent given *C* iff*:*
  - $P(A \wedge B \mid C) = P(A \mid C)\, P(B \mid C)$
  - Equivalently: $P(A \mid B \wedge C) = P(A \mid C)$ or $P(B \mid A \wedge C) = P(B \mid C)$
- Two random variables X and Y are conditionally independent given Z iff:
  - For all events A for X, B for Y, C for Z:
    A and B are conditionally independent given C
  - (discrete variables) Equivalently: $P(X{=}x, Y{=}y \mid Z{=}z) = P(X{=}x \mid Z{=}z) P(Y{=}y \mid Z{=}z)$
- *Note: independence does not imply conditional independence or vice versa*

Example 1

- **Conditional independence does not imply independence**
- On a random day,
  - A = event that Alice attends a lecture
  - B = event that Bob attends a lecture

$$P(A) = 3/7, \quad P(B) = 0.2$$

  - Given the event D that the day is either Mon, Wed or Fri

$$P(A|D) = 1, \quad P(B|D) = 0.7$$

  - Alice and Bob don't know each other, say $P(A \wedge B \mid D) = P(A|D)P(B|D)$
  - If Alice attends lecture, it's definitely M,W or F i.e. A,D are "duplicates"

$$P(B|A) = 0.7 \neq 0.2 = P(B)$$

  - A and B not independent, but are conditionally independent given D

Example 2

- **Independence does not imply conditional independence**
- Flip 2 coins.
- A = event coin 1 is heads
- B = event coin 2 is heads

$P(A|B) = P(A)$ A and B independent

- C = event exactly one coin was heads: C={HT,TH}

$$P(A|C) = \frac{1}{2}, \quad P(B|C) = \frac{1}{2}$$

$$P(A \wedge B \mid C) = 0 \neq P(A|C)P(B|C)$$

- A and B not conditionally independent given C

Expectation

- Denotes the expected value or mean value of a random variable X
- Discrete

$$E[X] = \sum_x x \;\; p(x)$$

- Continuous

$$E[X] = \int_x x \;\; p(x)dx$$

- Expectation of a function

$$E[aX + b] = a \;\; E[X] + b$$

$$E[h(X)] = \sum_x h(x) \;\; p(x)$$

$$E[h(X)] = \int_x h(x)p(x)dx$$

Called the “law of the unconscious statistician” (seriously)

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- X = {0, 1, 2, 3}
- Sample space: HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
- **X=0** (TTT), **X=1** (HTT, THT, TTH), **X=2** (HHT, HTH, THH), **X=3** (HHH)
- P(**X=0**) = 1/8; P(**X=1**) = 3/8; P(**X=2**) = 3/8; P(**X=3**) = 1/8
- What is the expected value of X, E[X]?

$$
\begin{aligned}
E[X] &= (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8}) \\
&= \frac{3}{2}
\end{aligned}
$$

Variance

- Denotes the squared deviation of X from its mean

- Variance

$$Var(X) = E[(X - E[X])^2]$$
$$= E[X^2] - (E[X])^2$$

- Standard deviation

$$\sigma = \sqrt{Var(X)}$$

- Variance of a function

$$Var(aX + b) = a^2 \quad Var(X)$$

$$Var(h(X)) = \sum_x (h(x) - E[h(X)])^2 \quad p(x)$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.

- X = {0, 1, 2, 3}

$$E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})$$
$$= \frac{3}{2}$$

- What is the variance of X, Var(X)?

$$\begin{aligned} Var(X) &= \left(\left[0 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) + \left(\left[1 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[2 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[3 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) \\ &= \left(\frac{9}{4} \cdot \frac{1}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{9}{4} \cdot \frac{1}{8}\right) \\ &= \frac{3}{4} \end{aligned}$$

Common distributions

- Bernoulli
- Binomial
- Multinomial
- Normal

Bernoulli

- Binary variable $X \in \{0, 1\}$ that takes the value of 1 with probability $p \in [0,1]$
  - E.g., Outcome of a fair coin toss is Bernoulli with p=0.5. Here x=1 means that the coin landed heads up, x=0 means the the coin landed tails up

$$P(x) = p^x(1-p)^{1-x}$$

$$E[X] = 1(p) + 0(1-p) = p$$

$$\begin{aligned} Var(X) &= E[X^2] - (E[X])^2 \\ &= 1^2(p) + 0^2(1-p) - p^2 \\ &= p(1-p) \end{aligned}$$

Binomial

- Describes the number of successful outcomes in *n* independent Bernoulli(*p*) trials

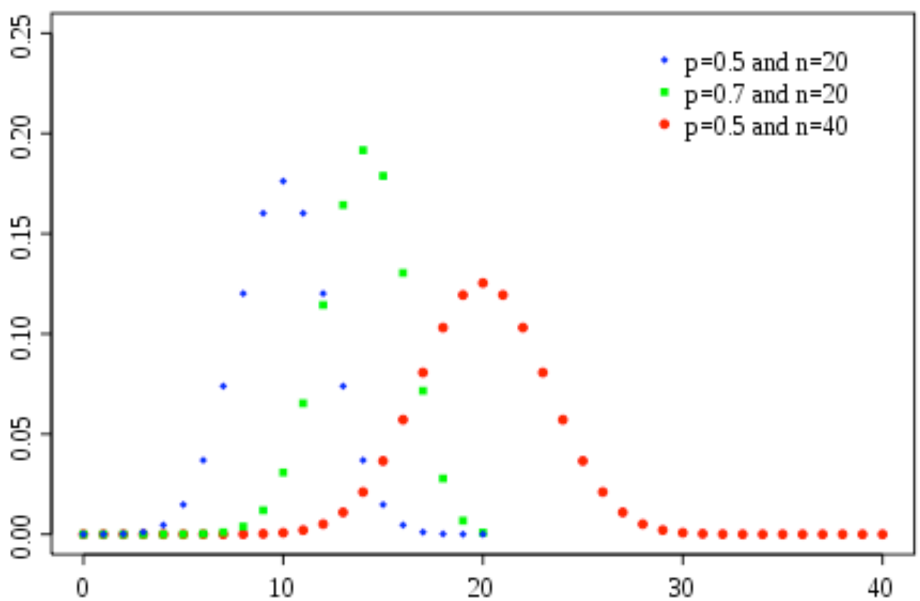
$$X \in \{0, 1, \ldots, n\}, \quad p \in [0, 1]$$

  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with *n*=10 and *p*=0.5. Here *x* is the number of heads.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E[X] = np$$

$$Var[X] = np(1-p)$$



Multinomial

- Generalization of binomial to *k* possible outcomes; outcome *i* has probability p_i of occurring; x_i is the number of times the *i*-th outcome occurs in *n* trials
  - E.g., Number of {outs, singles, doubles, triples, homeruns} in a sequence of n=10 times at bat is Multinomial. Here k=5, x_1 is the number of "outs", p_1 is the probability of "out", …, x_5 is the number of "homeruns", p_5 is the probability of "homerun".

$$x_i \in \{0,1,\ldots,n\}, \quad p_i \in [0,1], \quad \sum_{i=1}^{k} x_i = n, \quad \sum_{i=1}^{k} p_i = 1$$

$$P(x_1, \ldots x_k) = \binom{n}{x_1, \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

$$E[X_i] = np_i \qquad Var(X_i) = np_i(1 - p_i)$$

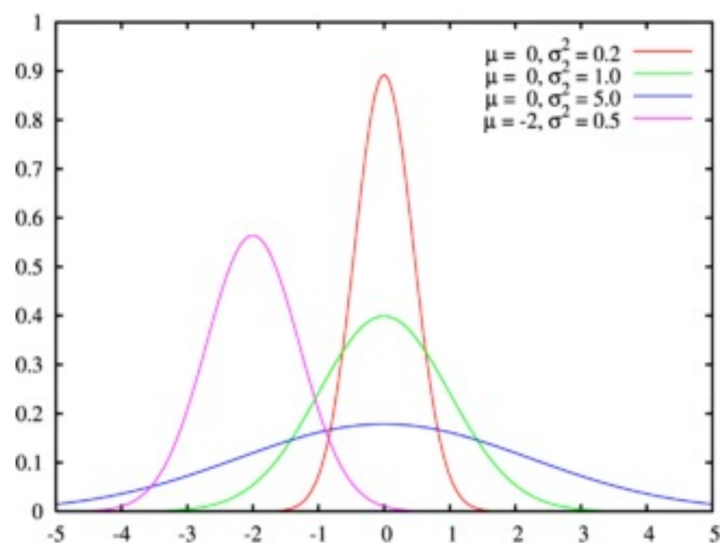
Normal (Gaussian)

- Important distribution gives well-known bell shape
- Central limit theorem:
  - Distribution of $\sqrt{n}$ times the average of n independent zero-mean samples becomes normally distributed as $n \rightarrow \infty$, regardless of the distribution of the underlying population

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$



Likelihood function

- A random variable $\underline{x}$ has **parameters** θ and probability $P(\underline{x};\theta)$

  e.g., Bernoulli: $\theta = p$, $P(x;\theta) = p^x(1-p)^{1-x}$

  multinomial: $\theta = (p_1,\ldots,p_k)$, $P(\underline{x};\theta) = \binom{n}{x_1,\ldots,x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$

- Assume we have *n* **independent** samples $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$

- Define the dataset $D = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$

- The likelihood function represents the probability of the dataset *D* as a function of the model parameters θ

$$L(D;\theta) = P(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n;\theta) = \prod_{i=1}^{n} P(\underline{x}_i;\theta)$$

by independence

Likelihood function

- The likelihood function represents the probability of the dataset *D* as a function of the model parameters θ

$$L(D;\theta) = P(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n; \theta) = \prod_{i=1}^{n} P(\underline{x}_i; \theta)$$

- Gives **relative probability of data given a parameter**
- We can compare two values θ and θ' by comparing their likelihoods
- We say that θ is better for explaining the dataset *D* than θ' if

$$L(D;\theta) > L(D;\theta')$$

Maximum likelihood estimation (MLE)

- Most widely used method of parameter estimation
- **Intuition:** a θ with higher likelihood explains better the data
- "Learn" the best parameters θ that maximizes likelihood:

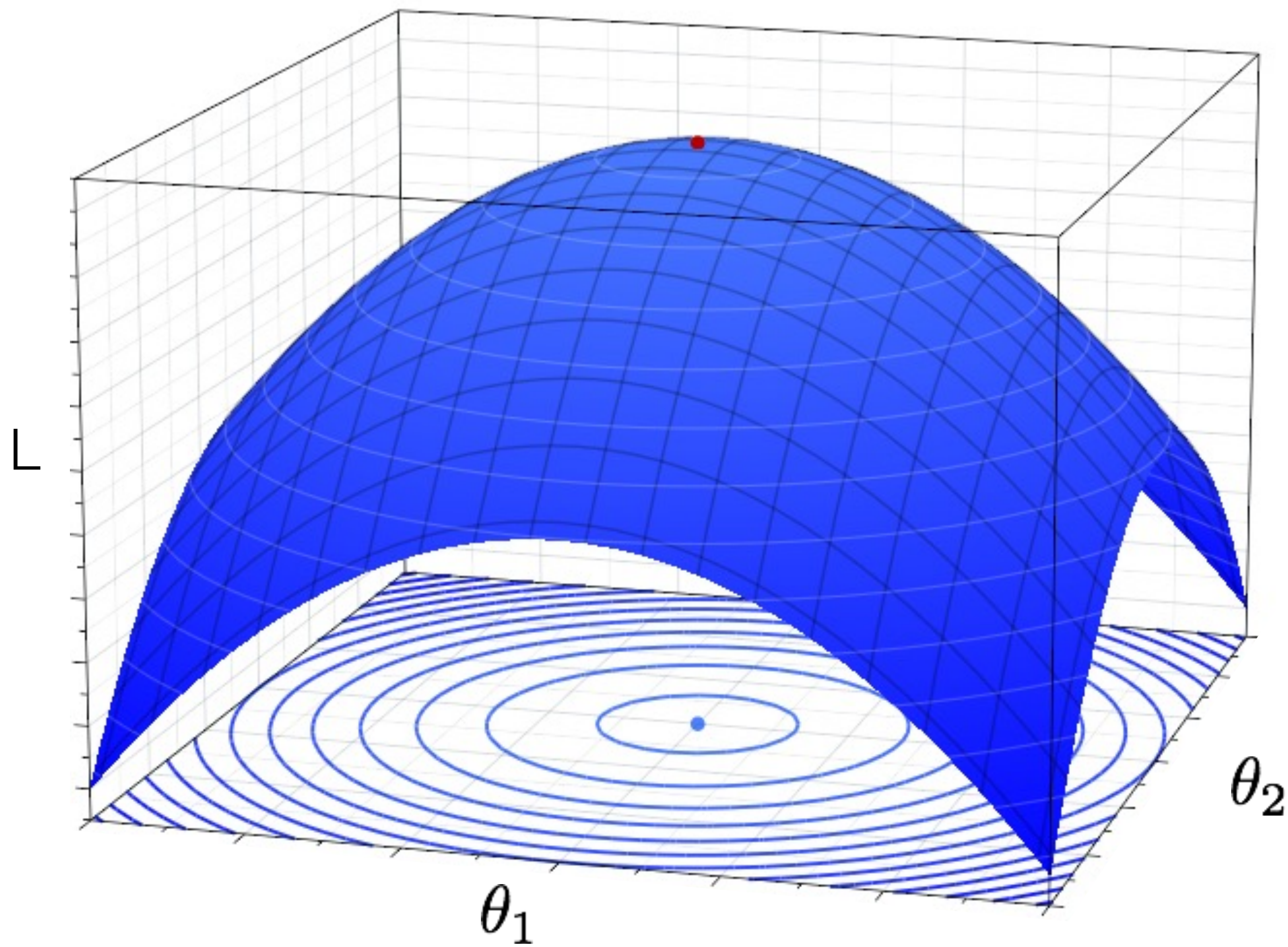
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; L(D;\theta)$$

- Often easier to work with log-likelihood:

$$l(D;\theta) = \log L(D;\theta) = \log \prod_{i=1}^{n} P(\underline{x}_i;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; l(D;\theta)$$

Likelihood surface



If the log-likelihood surface is concave, we can often determine the parameters that maximize the function analytically

Maximum Likelihood Estimation (MLE) for Bernoulli

- For a Bernoulli r.v. $x_i \in \{0,1\}$, $\theta = p$, $P(x_i;\theta) = p^{x_i}(1-p)^{1-x_i}$
- Clearly: $\log P(x_i;\theta) = x_i \log p + (1-x_i)\log(1-p)$
- The **log-likelihood function** is:

$$l(D;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i;\theta)$$
$$= \sum_{i=1}^{n} \big(x_i \log p + (1-x_i)\log(1-p) \big)$$
$$= \left(\sum_{i=1}^{n} x_i \right) \log p + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p)$$

- Recall that the **MLE** is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; l(D;\theta)$$

Maximum Likelihood Estimation (MLE) for Bernoulli

- For a Bernoulli r.v. $x_i \in \{0,1\}$, $\theta = p$, $P(x_i;\theta) = p^{x_i}(1-p)^{1-x_i}$
- The **log-likelihood function** is:

$$l(D;\theta) = \left(\sum_{i=1}^{n} x_i\right)\log p + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p)$$

- Recall that the **MLE** is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(D;\theta)$$

- We can maximize $l(D;\theta)$ by taking derivative equal to zero:

$$\frac{\partial l(D;\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \qquad \text{then} \qquad \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The MLE $\hat{\theta} = \hat{p}$ is the **proportion of ones in the dataset**. This is intuitive since the parameter $\theta = p = \mathrm{E}[X]$ is the **expected proportion of ones**.

Maximum Likelihood Estimation (MLE) for Bernoulli

```
import numpy as np
def example_bernoulli(n):
  z = np.random.randint(0,2,n)
  return 1.0/n * np.sum(z)

>>> example_bernoulli(10)
0.8
>>> example_bernoulli(100)
0.44
>>> example_bernoulli(10000)
0.5138
```

Returns n random integers >= 0 and < 2, each value with equal probability.
In this case (0 or 1) then p = 0.5 in the Bernoulli distribution

Computes average

From the terminal, use your Career account to start a session:

    ssh username@data.cs.purdue.edu

From the terminal:

    python