CS464
Chapter 2:
Review of Probability

(based on the slides provided by Öznur Taştan and Mehmet Koyutürk)

Uncertainty

- There are certain statements that you know to be *true.*

- There are certain statements that you know to be *false.*

- But with the majority of the statements, you do not know whether they are true or false; for you, *these statements are uncertain*
  - In machine learning, this is almost always the case: we are trying to identify patterns that indicate labels, but the relationship between the patterns and labels is usually uncertain

Sample Space

- **Sample space Ω :** The set of all possible outcomes of a random experiment.

  - If the experiment concerns in determination of an email being spam or ham $\Omega = \{\mathrm{s}, \mathrm{h}\}$

  - For tossing a dice, $\Omega = \{1,2,3,4,5,6\}$.

  – If outcome of an experiment is the order of n documents when the word “machine learning” is queried in the web search engine

$$\Omega = \{\text{all n! permutations of } (1, \ldots, \mathrm{n})\}$$

Event

- **An event** is any subset of sample space Ω.
- The **space of events** is a subset of the power set of Ω.

  - For email classification, *A* can be the sevent that the e--mail is spam
    - $A = \{s\}$

  - For tossing a dice, A = {2,4,6} is the “number is even” event.

  - For document ranking, *A* can be the event that Document 5 is ranked first
    - $A =$ {all permutations of *n* documents such that Document 5 comes first}

Basic Concepts (from Set Theory)

- The **union of two events** A and B, $A \cup B$, is the event consisting of all outcomes that are either in A or in B or in both events.

- The **complement of an event** A, A^c, is the set of all outcomes in Ω that are not in A.

- The **intersection of two events** A and B, $A \cap B$, is the event consisting of all outcomes that are in **both events**.

- When two events A and B have no outcomes in common $A \cap B = \emptyset$, they are said to be **mutually exclusive**, or **disjoint**, events.

Axioms of Probability

- For any event $X \in \Omega$, $0 \leq P(X) \leq 1$

- $P(\Omega) = 1$

- For any sequence of **mutually exclusive events** that is $X_Q \cap X_R = \emptyset$ for all $i \neq j$

$$P\left(\cup_Q X_Q\right) = \sum_Q P(X_Q)$$

Probability of Union of Two Events

Given two events A and B, the **union** of two events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If events A and B are **mutually exclusive**, then

$$P(A \cup B) = P(A) + P(B)$$

Random Variable

- A random variable, *X*, is a real--valued function defined on a set of possible outcomes (i.e., the sample space Ω)

$$X: \Omega \rightarrow \mathbb{R}$$

- Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes

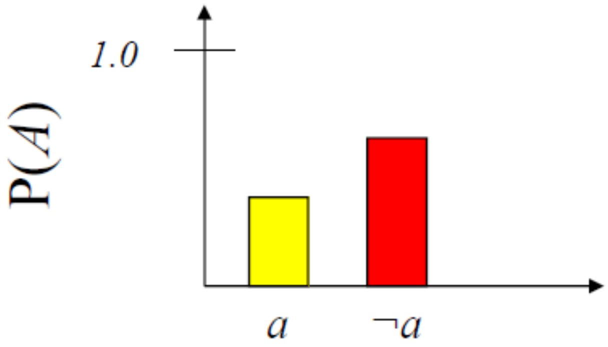
  – Examples: spam/ham, cancer/noncancer, dice counts, etc.

- **Continuous** random variables have an infinite continuum of possible values

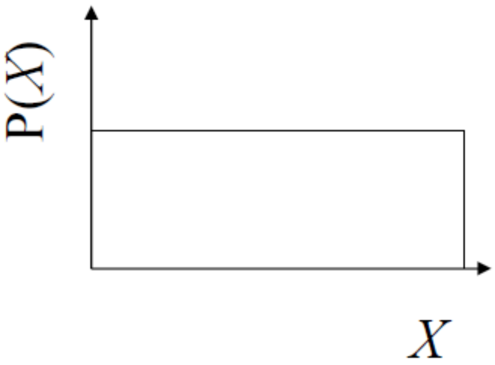
  – Examples: blood pressure, weight

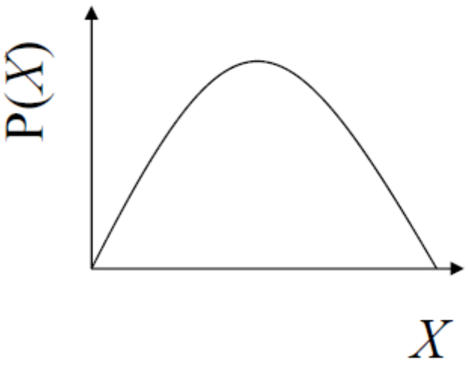
Probability Mass/Density Function

- Discrete random variables have probability mass distributions



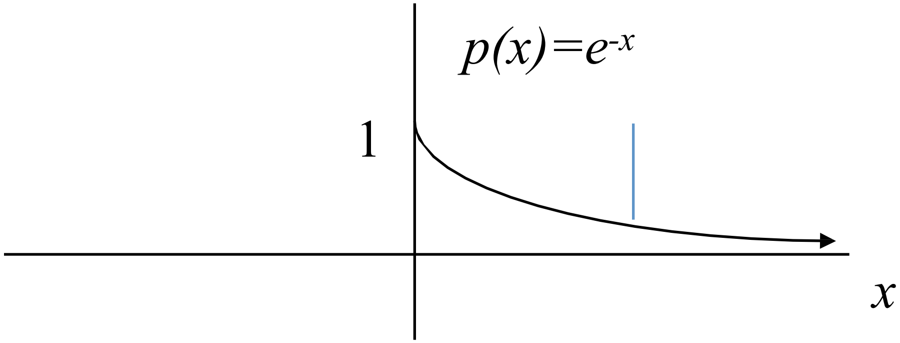
- Continuous random variables have probability density functions





Probability Density Function (pdf)

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.



- The probability that *x* is any exact particular value (such as 1.9976) is 0

- We can only assign probabilities to possible ranges of x.

A Bernoulli Trial

- A Bernoulli trial is a trial with a binary outcome, for which the probability that the outcome is 1 equals θ

Bernoulli (θ):

$$\mathbf{P}(X = 1) = \theta$$

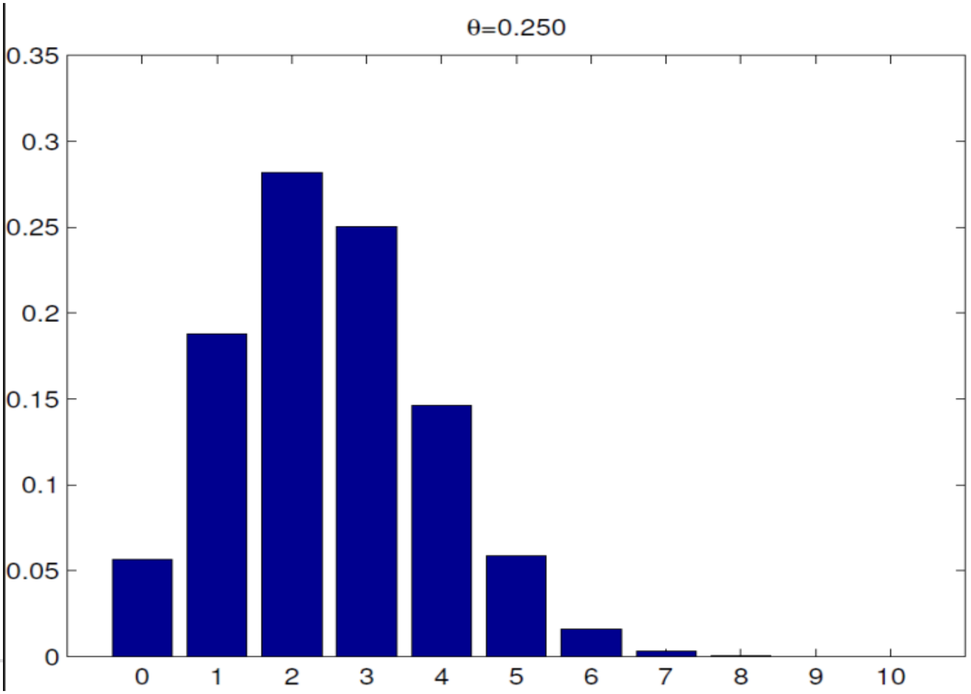
$$\mathbf{P}(X = 0) = 1 - \theta$$

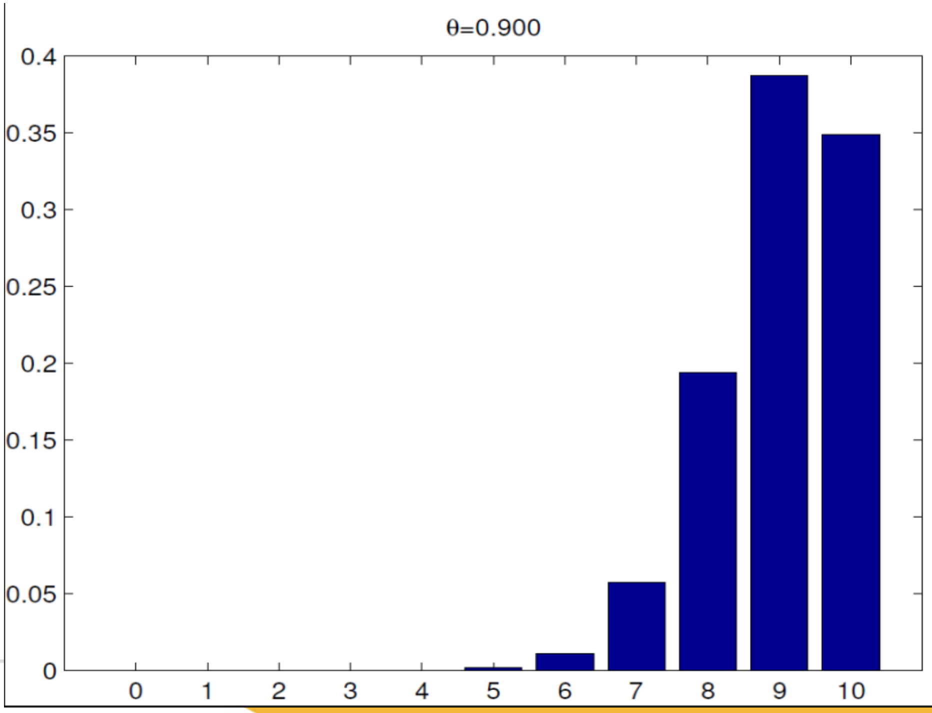
Binomial Distribution

- Number k of successes from n independent Bernoulli trials.

Binomial(n, θ):

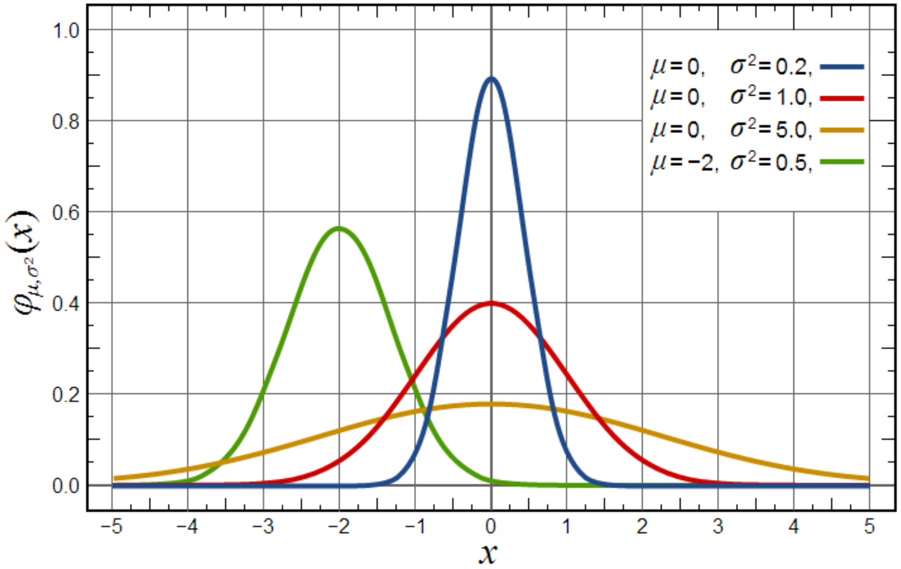
$$\mathbf{P}\left(X=k\right)=\binom{n}{k}\theta^{k}(1-\theta)^{n-k}$$





Gaussian (Normal) Distribution

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



μ = mean of distribution

σ^2 = variance of distribution

x is a continuous variable ($-\infty \leq x \leq \infty$)

- Many distributions in nature can be modeled using normal distribution
  - Heights of people, scores received in an exam, environmental noise in audio signals…

Expected Value and Variance

- All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).

Expected Value

- Expected value is just the average or mean (μ) of random variable *X*.

- It’s a “weighted average” because more frequent values of X are weighted more in the average

Expected Value

Discrete case:

$$E(X) = \sum_{\text{all x}} x_i p(x_i)$$

Continuous case:

$$E(X) = \int_{\text{all x}} x_i p(x_i) dx$$

Expected Value

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

$$\sum_{i=1}^{5} x_i p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Prior Probability

- Prior probability is the probability of event A in the absence of any other information

- If we get new information that affects *A*, we can reason with the conditional probability of *A* given the new information.

Joint Probability

- Often, we need to consider the relationship between two or more events:

P(email being a spam AND email containing the word free)

- Joint probability distributions allow us to reason about the relationship between multiple events

Joint Probability Distributions

- A joint probability distribution over a set of random variables $X_1, X_2, \ldots, X_\mathrm{n}$ specifies a probability for each assignment:

$$\mathbf{P}\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n\right)$$

- Let's use short hand notation:

$$\mathbf{P}\left(x_1, x_2, ..., x_n\right)$$

Full Joint Probability Distribution

- Suppose you have the complete set of random variables used to describe the world.

- A joint probability distribution that covers this complete set is called the full joint probability distribution.

- Is a complete specification of one’s uncertainty about the world in question. Very powerful.

Joint Probability Distributions

- Must obey:

$$V x \;\; P(x_{\mathrm{a}},\ x_{\mathrm{f}},..,x_{\mathrm{i}}) \geq 0$$

$$\sum\nolimits_{(\mathrm{k_l},\ \mathrm{k_m},..,\mathrm{k_n})} P(x_{\mathrm{a}},\ x_{\mathrm{f}},..,x_{\mathrm{i}}) = 1$$

- Size of the joint distribution with *n* variables and each having *d* possible outcomes?

- **For all, except very small distributions, impractical to write out.**

P(T, W)

T	W	P
hot	sun	0.20
hot	cloudy	0.05
hot	rain	0.05
warm	sun	0.20
warm	cloudy	0.10
warm	rain	0.05
cold	sun	0.05
cold	cloudy	0.10
cold	rain	0.20

Marginalization

- In many situations, we may only care about the probability distribution of one variable (e.g. *P(sunny)),* and do not care about the probability of the other variables

- The marginal distribution refers to the probability distribution of a random variable on its own

- Marginalization allows us to eliminate the dependence of the probability distribution on a given variable by "integrating it out"

$$\mathbf{P}\left(X_1 = x_1\right) = \sum_{y_i} \mathbf{P}\left(X = x_i, Y = y_i\right)$$

Marginal Distribution

P(T, W)

T	W	P
hot	sun	0.20
hot	cloudy	0.05
hot	rain	0.05
warm	sun	0.20
warm	cloudy	0.10
warm	rain	0.05
cold	sun	0.05
cold	cloudy	0.10
cold	rain	0.20

Marginalizing over *T* to get ***P(W)***

P(W)

W	P
sun	0.45
rain	0.30
cloudy	0.25

Marginal Distribution

P(T, W)

T	W	P
hot	sun	0.20
hot	cloudy	0.05
hot	rain	0.05
warm	sun	0.20
warm	cloudy	0.10
warm	rain	0.05
cold	sun	0.05
cold	cloudy	0.10
cold	rain	0.20

Marginalizing over ***W*** to get ***P(T)***

P(T)

T	P
hot	0.30
cold	0.35
warm	0.35

Conditional Probability

- The ***conditional probability*** of an event *A* is the probability that the event will occur given the knowledge that an event *B* has already occurred.

  P(an email being spam | email contains "free")

- When $P(B) > 0$ the conditional probability of any event A given B is defined as:

$$\mathbf{P}(A \mid B) \triangleq \frac{\mathbf{P}(A,B)}{\mathbf{P}(B)}$$

Product Rule

$$\mathbf{P}\left(A \mid B\right) = \frac{\mathbf{P}(A,B)}{\mathbf{P}(B)} \Rightarrow \mathbf{P}\left(A, B\right) = \mathbf{P}\left(A \mid B\right) \mathbf{P}\left(B\right)$$

Bayes' Rule

$$\mathbf{P}(Y) = \mathbf{P}(Y \mid X)\,\mathbf{P}(X) + \mathbf{P}(Y \mid X^c)\,\mathbf{P}(X^c)$$

which simply states that the probability of event *Y* is the sum of the conditional probabilities of event *Y* given that event *X* has or has not occurred.

Then we can write the conditional probabilities:

$$\mathbf{P}(X \mid Y) = \frac{\mathbf{P}(X,Y)}{\mathbf{P}(Y)} = \frac{\mathbf{P}(X)\mathbf{P}(Y \mid X)}{\mathbf{P}(Y \mid X)\mathbf{P}(X)+\mathbf{P}(Y \mid X^c)\mathbf{P}(X^c)}$$

Conditional Distribution

What is the conditional distribution $P(W \mid T)$

$P(W \mid T = hot)$

$P(W \mid T = cold)$

$P(W \mid T = warm)$

T	W	P
hot	sun	0.20
hot	cloudy	0.05
hot	rain	0.05
warm	sun	0.20
warm	cloudy	0.10
warm	rain	0.05
cold	sun	0.05
cold	cloudy	0.10
cold	rain	0.20

Conditional Probability

$P(W \mid T = hot)$

W	P
sun	2/3
cloudy	1/6
rain	1/6

$P(W \mid T = cold)$

W	P
sun	5/35
cloudy	10/35
rain	20/35

$P(W \mid T = warm)$

W	P
sun	20/35
cloudy	10/35
rain	5/35

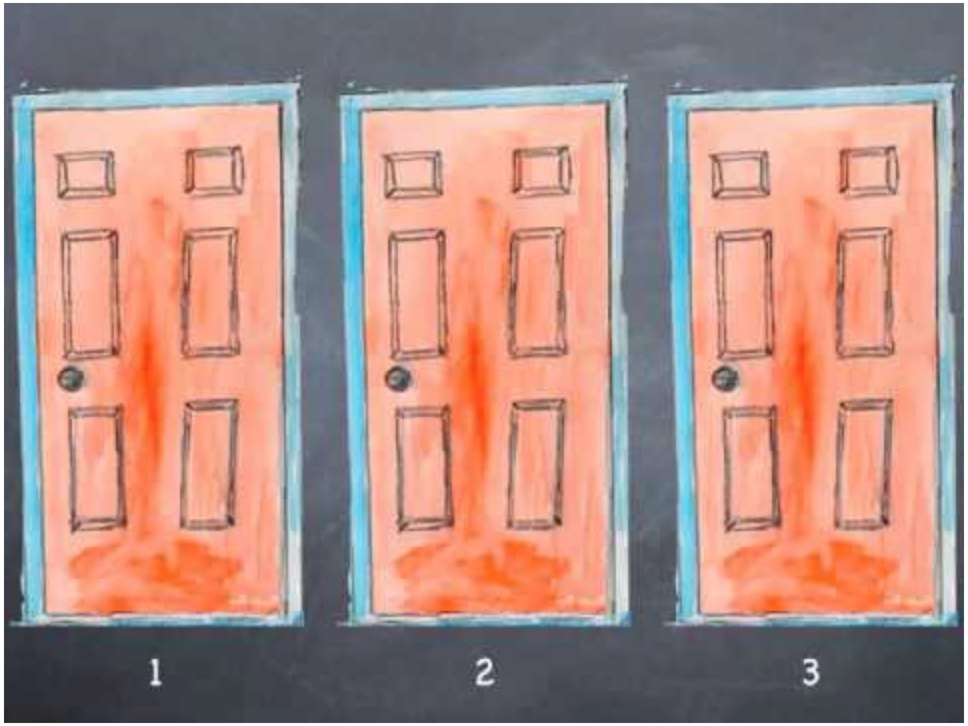
Chain Rule

$$P(X_1, X_{2,} \dots, X_n) = P(X_1)\, P(X_2 | X_1) \dots P(X_n \mid X_{1,} \dots, X_{n-1})$$

- Used to evaluate the joint probability of some random variables, and is especially useful when there are (conditional) independence across variables.
- Notice there is a choice in the **order** we unravel the random variables when applying the Chain Rule; picking the right order can often make evaluating the probability much easier.
- Example: p(A,B,C,D)=p(A)p(B|A)p(C|A,B)p(D|A,B,C)

  =p(D)p(B|D)p(A|B,D)p(C|A,B,D)

The Monty Hall Problem



Independence

- Two events are called independent $X \perp Y$ if and only if

$$X \perp\!\!\!\perp Y \Leftrightarrow \mathbf{P}\left(X, Y\right) = \mathbf{P}\left(X\right)\mathbf{P}\left(Y\right)$$

Or equivalently,

$$\mathbf{P}\left(X \,|\, Y\right) = \mathbf{P}\left(X\right)$$

Conditional Independence

- X and Y are conditionally independent given Z iff the conditional probability of the joint variable can be written as product of conditional probabilities:

$$X \perp Y|Z \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- The following slides are sloppy and include “intentional” mistakes on the slides

Marginals, Conditionals

Given two discrete random variables X and Y

X takes values in $\{x_1, \dots, x_m\}$

Y takes values in $\{y_1, \dots, y_n\}$

$$P(Y = y_j) = \sum_j P(X = x_i, Y = y_j)$$

$$P(Y = y_j) = \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

Marginals, Conditionals

Given two discrete random variables X and Y

X takes values in $\{x_1, \dots, x_m\}$

Y takes values in $\{y_1, \dots, y_n\}$

$$P(Y = y_j) = \sum_j P(X = x_i, Y = y_j)$$

$$P(Y = y_j) = \sum_j P(X = x_i \mid Y = y_j) P(Y = y_j)$$

Marginals, Conditionals

Given two discrete random variables X and Y

X takes values in $\{x_1, \dots, x_m\}$

Y takes values in $\{y_1, \dots, y_n\}$

$P(X = x_i)$

$$\cancel{P(Y = y_j)} = \sum_j P(X = x_i, Y = y_j)$$

$$\cancel{P(Y = y_j)} = \sum_j P(X = x_i \mid Y = y_j) P(Y = y_j)$$

$P(X = x_i)$

Law of Total Probability

Given two discrete random variables X and Y

X takes values in $\{x_1, \ldots, x_m\}$

Y takes values in $\{y_1, \ldots, y_n\}$

$$\boxed{P(X = x_i) = \sum_j P(X = x_i, Y = y_j)}$$

$$\boxed{P(X = x_i) = \sum_j P(X = x_i \mid Y = y_j) P(Y = y_j)}$$

Marginals, Conditionals

Given two discrete random variables X and Y

Joint probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

$$= \sum_j \underline{P(X = x_i \mid Y = y_j)P(Y = y_j)}$$

Marginal probability

Conditional probability of X conditioned on Y

Marginals, Conditionals

Given two discrete random variables X and Y

Joint probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

$$= \sum_j \underline{P(X = x_i | Y = y_j)P(Y = y_j)}$$

Marginal probability

Conditional probability of X conditioned on Y

Marginals, Conditionals

Given two discrete random variables X and Y

Joint probability of X,Y

$$P(X = x_i) = \sum_j \overline{P(X = x_i, Y = y_j)}$$

$$= \sum_j \underline{P(X = x_i \mid Y = y_j)}P(Y = y_j)$$

Marginal probability

Conditional probability of X conditioned on Y

Marginal probability of Y

In a Strange World

Two discrete random variables *X* and *Y* take binary values

Joint probabilities

$$P(X = 0, Y = 1) = 0.2$$

$$P(X = 0, Y = 0) = 0.2$$

$$P(X = 1, Y = 0) = 0.5$$

$$P(X = 1, Y = 1) = 0.5$$

In a Strange World

Two discrete random variables *X* and *Y* take binary values

Joint probabilities

Should sum up to 1

$$P(X=0,Y=1)=\cancel{0.2}$$

$$P(X=0,Y=0)=\cancel{0.2}$$

$$P(X=1,Y=0)=\cancel{0.5}$$

$$P(X=1,Y=1)=\cancel{0.5}$$

The World Seems Fine

Two discrete random variables *X* and *Y* take binary values

Joint probabilities

$$P(X=0, Y=1)=0.2$$

$$P(X=0, Y=0)=0.2$$

$$P(X=1, Y=0)=0.3$$

$$P(X=1, Y=1)=0.3$$

What about the marginals?

Joint probabilities

$$P(X = 0, Y = 1) = 0.2$$

$$P(X = 0, Y = 0) = 0.2$$

$$P(X = 1, Y = 0) = 0.3$$

$$P(X = 1, Y = 1) = 0.3$$

Marginal probabilities

$$P(X = 0) = 0.2$$
$$P(X = 1) = 0.8$$

$$P(Y = 0) = 0.5$$
$$P(Y = 1) = 0.5$$

This is a Strange World

Joint probabilities

$$P(X=0,Y=1)=0.2$$

$$P(X=0,Y=0)=0.2$$

$$P(X=1,Y=0)=0.3$$

$$P(X=1,Y=1)=0.3$$

Marginal probabilities

~~$P(X=0)=0.2$~~

~~$P(X=1)=0.8$~~

$$P(Y=0)=0.5$$

$$P(Y=1)=0.5$$

World is in Order

Joint probabilities

Marginal probabilities

$$\begin{tikzcd}
P(X = 0, Y = 1) = 0.2 \arrow[r] & P(X = 0) = 0.4 \\
P(X = 0, Y = 0) = 0.2 \arrow[ur] & P(X = 1) = 0.6 \\
P(X = 1, Y = 0) = 0.3 \arrow[ur] & P(Y = 0) = 0.5 \\
P(X = 1, Y = 1) = 0.3 \arrow[uur] & P(Y = 1) = 0.5
\end{tikzcd}$$

$$P(X = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 1) = 0.4$$

Conditional Probabilities

What is the complementary event of $P(X=0|Y=1)$?

$P(X=1|Y=1)$ OR $P(X=0|Y=0)$

Conditional Probabilities

What is the complementary event of $P(X=0|Y=1)$?

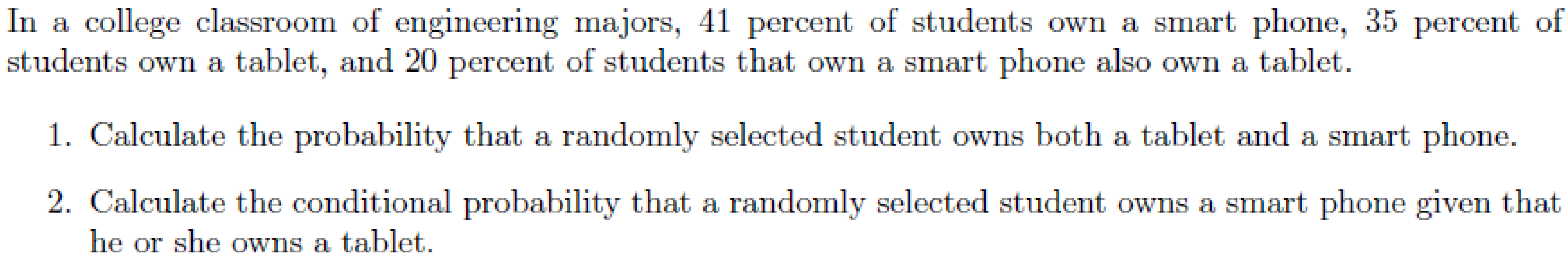
$P(X=1|Y=1)$

For your future reference

Slides with ***mistakes*** are marked with → !?
for your future reference

(Supposedly) Correct slides
are marked with ☑

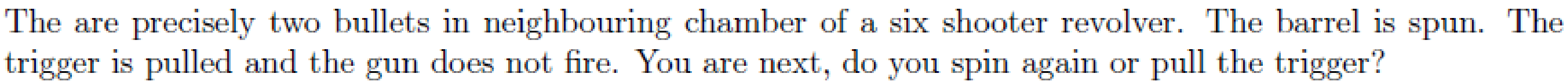
Problem 1: These kids are spoiled



In a college classroom of engineering majors, 41 percent of students own a smart phone, 35 percent of students own a tablet, and 20 percent of students that own a smart phone also own a tablet.

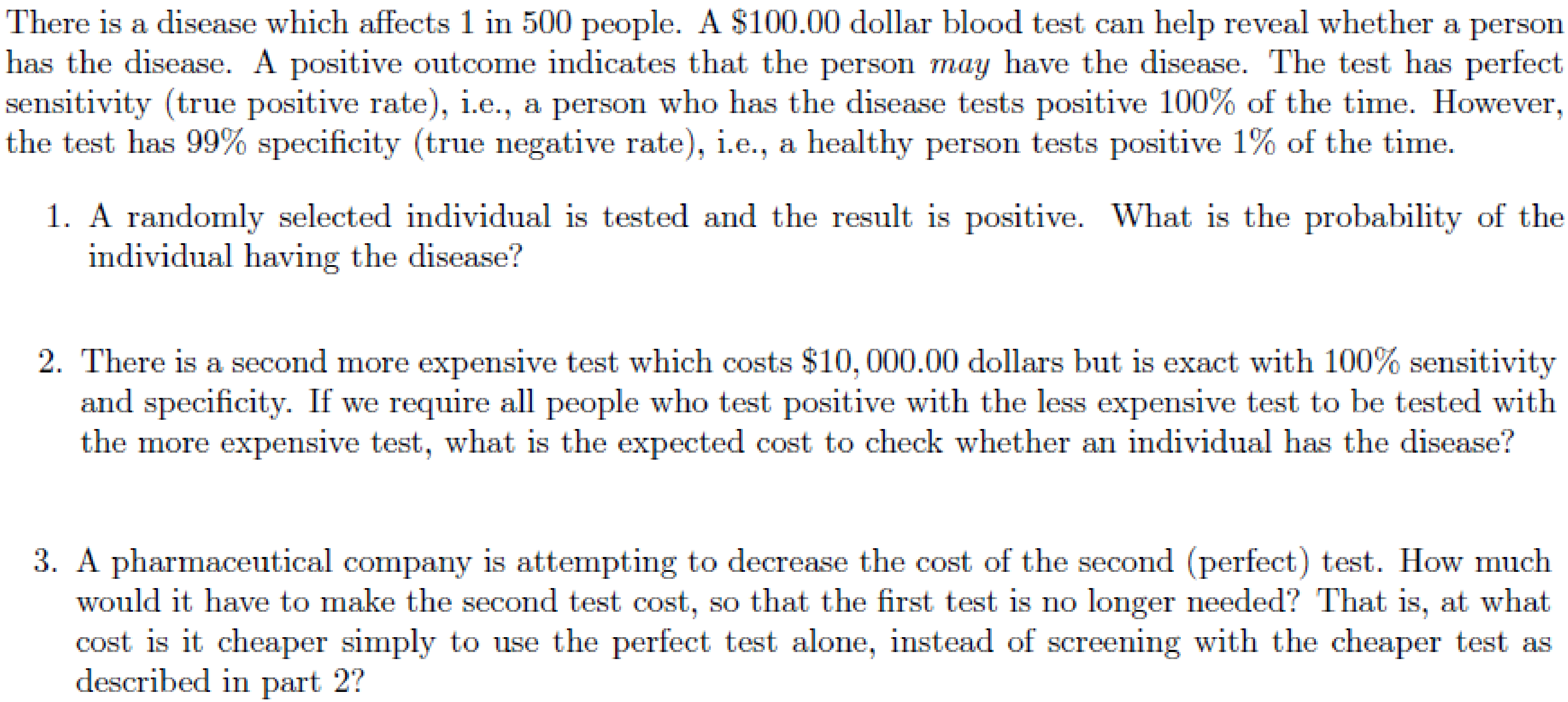
1. Calculate the probability that a randomly selected student owns both a tablet and a smart phone.
2. Calculate the conditional probability that a randomly selected student owns a smart phone given that he or she owns a tablet.

Problem 2: To pull or not to pull



The are precisely two bullets in neighbouring chamber of a six shooter revolver. The barrel is spun. The trigger is pulled and the gun does not fire. You are next, do you spin again or pull the trigger?

Problem 3: Optimizing medical tests



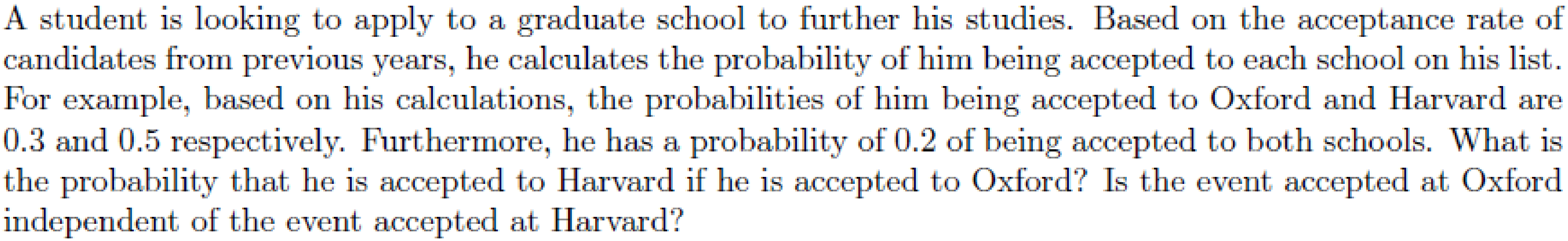
There is a disease which affects 1 in 500 people. A $100.00 dollar blood test can help reveal whether a person has the disease. A positive outcome indicates that the person *may* have the disease. The test has perfect sensitivity (true positive rate), i.e., a person who has the disease tests positive 100% of the time. However, the test has 99% specificity (true negative rate), i.e., a healthy person tests positive 1% of the time.

1. A randomly selected individual is tested and the result is positive. What is the probability of the individual having the disease?

2. There is a second more expensive test which costs $10,000.00 dollars but is exact with 100% sensitivity and specificity. If we require all people who test positive with the less expensive test to be tested with the more expensive test, what is the expected cost to check whether an individual has the disease?

3. A pharmaceutical company is attempting to decrease the cost of the second (perfect) test. How much would it have to make the second test cost, so that the first test is no longer needed? That is, at what cost is it cheaper simply to use the perfect test alone, instead of screening with the cheaper test as described in part 2?

Problem 4: Harvard vs. Oxford



A student is looking to apply to a graduate school to further his studies. Based on the acceptance rate of candidates from previous years, he calculates the probability of him being accepted to each school on his list. For example, based on his calculations, the probabilities of him being accepted to Oxford and Harvard are 0.3 and 0.5 respectively. Furthermore, he has a probability of 0.2 of being accepted to both schools. What is the probability that he is accepted to Harvard if he is accepted to Oxford? Is the event accepted at Oxford independent of the event accepted at Harvard?