CSE 312

Foundations of Computing II

Lecture 10: Naïve Bayes + Random Variables Intro

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Stefano Tessaro

tessaro@cs.washington.edu

Machine Learning

Used to derive decision rules from data, where no clear set of rules apply.

- *Is this the picture of a cat or of a dog?*
- *Should the car slow down?*
- *Is this e-mail spam?*
- *Which digit is in this picture?*
- *What is the translation of this text?*
- *Does this patient have disease X?*
- *Where are the faces on this picture?*
- *…*

Setting – Supervised Learning

Training data

labeled

Temp	Heart rate	Cough	BP	Sore throat	disease?
104	105	Y	85	Y	Y
97	90	N	90	N	N
..					..
101	100	Y	105	N	Y

Learner

Accuracy of classifier usually evaluated on test data != training data.

Temp	Heart rate	Cough	BP	Sore throat
102	99	N	95	Y

Classifier

Has disease / No disease

Today – Naïve Bayes Algorithm

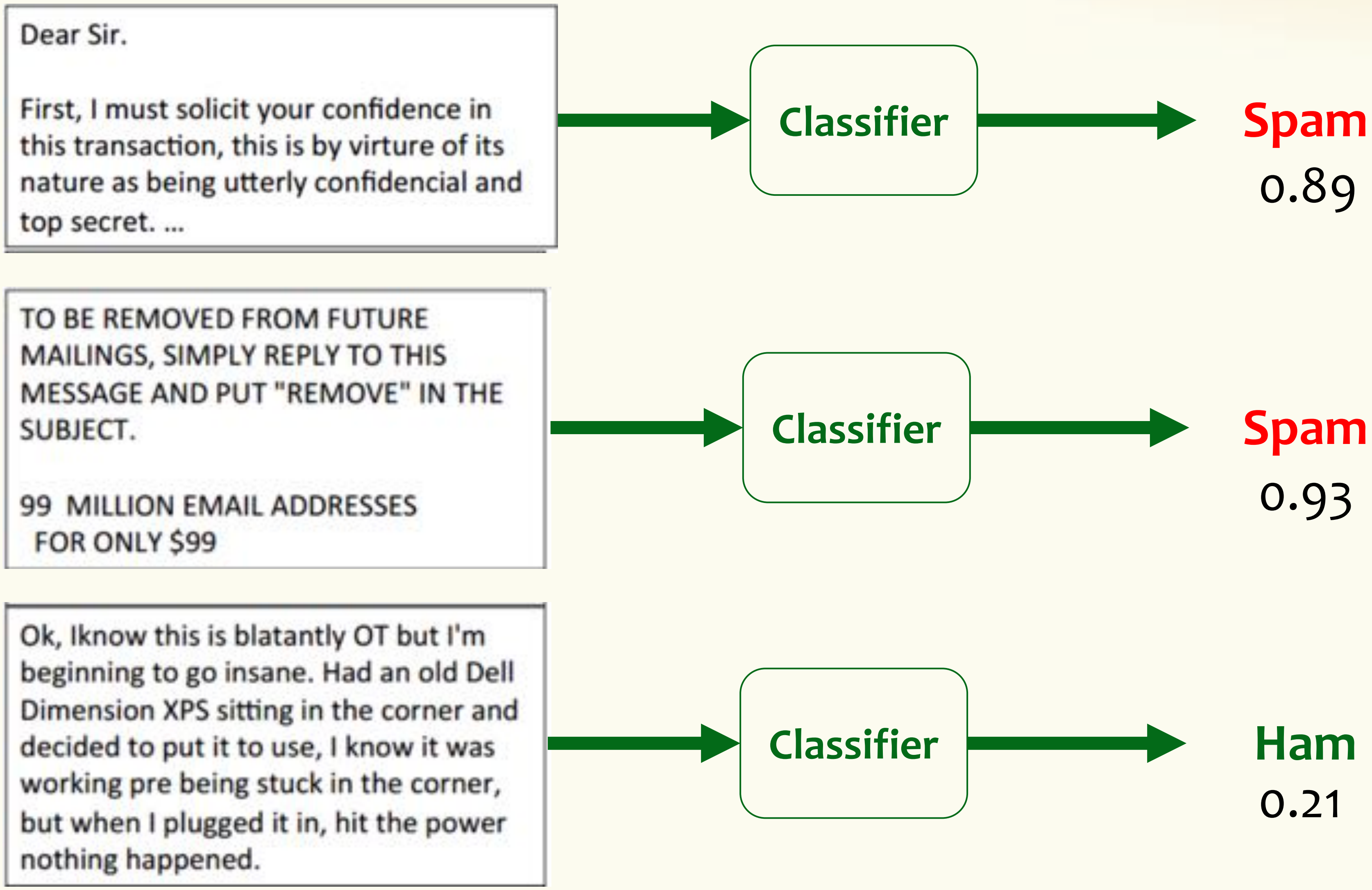
Classifier based on Bayes Rule.

Canonical application: **Spam filtering**

- used by Gmail, Bogofilter, DSPAM, SpamBayes, ASSP, CRM114.Mozilla Thunderbird, Mailwasher, SpamAssassin

But also used for:

- Sentiment analysis in text
- Medical diagnosis
- Market predictions
- …


Dear Sir.
First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. ...
Classifier
Spam
0.89
TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.
99 MILLION EMAIL ADDRESSES FOR ONLY $99
Classifier
Spam
0.93
Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
Classifier
Ham
0.21

Naïve Bayes – Approach

To start our task, we are given **training data:**

- Several e-mails, labeled **spam** / **ham**
  – This needs to be done by someone, often by hand
- Possible features give away whether e-mail is spam or ham
  – words in body, subject line, sender, message header, time sent
- Here, simplification: We only look at **words** in document!

E-mails as word collections

E-mail

SUBJECT: Top Secret Business Venture

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret…

Set of words in document

{top, secret, business, venture, dear, sir, first, I, must, solicit, your, confidence, in, this, transaction, is, by, virture, of, its, nature, as, being, utterly, confidencial, and}

Naïve Bayes

Given document with set of words $\{w_1 \dots, w_n\}$

Goal: Classifier outputs (estimation of) $\mathbb{P}(\text{spam} \mid \{w_1 \dots, w_n\})$

How to compute?

Idea: Use Bayes Rule

$$\mathbb{P}(\text{spam} \mid \{w_1 \dots, w_n\}) = \frac{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam})}{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) + \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham})}$$

How do we compute the individual values? Estimate from training data!

Naïve Bayes – Estimating Parameters

$$\mathbb{P}(\text{spam} \mid \{w_1 \dots, w_n\}) = \frac{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam})}{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) + \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham})}$$

Estimate from training data!

$$\mathbb{P}(\text{spam}) = \frac{\#\text{ spam emails in training data}}{\#\text{emails in training data}} \qquad \mathbb{P}(\text{ham}) = 1 - \mathbb{P}(\text{spam})$$

$$\left.\begin{array}{l} \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) = ? \\ \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham}) = ? \end{array}\right\}$$

Problem: We likely do not have a document with words $\{w_1 \dots, w_n\}$ in training data!

Naïve Bayes – Assumption

Definition. $\mathcal{A}$ and $\mathcal{B}$ **independent conditioned on** $\mathcal{C}$

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B} | \mathcal{C}) = \mathbb{P}(\mathcal{A}|\mathcal{C}) \cdot \mathbb{P}(\mathcal{B}|\mathcal{C})$$

Conditional independence, i.e., conditioned on spam / ham, occurrences of individual words are independent.

$$\mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid \text{spam})$$

$$\mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham}) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid \text{ham})$$

Note: This is a strong assumption (hence, "naïve") – works just well in practice.

Naïve Bayes – Estimating Parameters

$$\mathbb{P}(\text{spam} \mid \{w_1 \dots, w_n\}) = \frac{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam})}{\mathbb{P}(\text{spam}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) + \mathbb{P}(\text{ham}) \cdot \mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham})}$$

$$\mathbb{P}(\text{spam}) = \frac{\#\text{ spam emails in training data}}{\#\text{emails in training data}}$$

$$\mathbb{P}(\text{ham}) = 1 - \mathbb{P}(\text{spam})$$

$$\mathbb{P}(w_i \mid \text{spam}) = \frac{\#\text{ spam emails in TD with } w_i}{\#\text{ spam emails}}$$

$$\mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid \text{spam})$$

$$\mathbb{P}(w_i \mid \text{ham}) = \frac{\#\text{ ham emails in TD with } w_i}{\#\text{ ham emails}}$$

$$\mathbb{P}(\{w_1, \dots, w_n\} \mid \text{ham}) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid \text{ham})$$

Does this work?

Imagine no spam e-mail in training set contains the word "Hullaballoo", but one ham e-mails contains it.

What is the problem?

$$\mathbb{P}(\text{"Hullaballoo"} \mid \text{spam}) = \frac{\#\text{ spam emails in TD with "Hullaballoo"}}{\#\text{ spam emails}} = 0$$

Recall: $\mathbb{P}(\{w_1, \dots, w_n\} \mid \text{spam}) = \prod_{i=1}^{n} \mathbb{P}(w_i \mid \text{spam})$

SUBJECT: Get out of debt! Cheap prescription pills! Earn fast cash using this one weird trick! Meet singles near you and get preapproved for a low interest credit card! Hullaballoo

$$\mathbb{P}(\text{spam} \mid \{w_1 \dots, w_n\}) = 0$$

Laplace Smoothing

Idea: Add two dummy spam e-mails. One contains every word appearing in training set, one contains none!

$$\mathbb{P}(w_i \mid \text{spam}) = \frac{\#\text{ spam emails in TD with } w_i + 1}{\#\text{ spam emails} + 2}$$

Project – Try it out yourself!

Project will be posted on Friday night

- Due on Nov 6 (tentative)
- Optional – but if submitted, will give small homework incentive.
- More information on Friday on edstem.

Next: Random Variables

Random Variables – First encounter

Often: We want to **capture quantitative properties** of the outcome of a random experiment, e.g.:

- *What is the total of two dice rolls?*
- *What is the number of coin tosses needed to see the first head?*
- *What is the number of heads among 20 coin tosses?*

Random Variables

Definition. A **random variable (RV)** for a probability space $(\Omega, \mathbb{P})$ is a function $X: \Omega \to \mathbb{R}$.*

Example. Throwing two dice $\Omega = \{(i, j) \mid i, j \in [6]\}$ $\mathbb{P}((i, j)) = \frac{1}{36}$.

$$\left.\begin{aligned} X(i, j) &= i + j \\ Y(i, j) &= i \cdot j \\ Z(i, j) &= i \end{aligned}\right\} \text{Random variables!}$$

* random variables outputting values from a non-numeric set can also be defined.

Random Variables

Definition. For a RV X, we define the event

$$\{X = x\} = \{\omega \in \Omega \mid X(\omega) = x\}$$

We write $\mathbb{P}(X = x) = \mathbb{P}(\{X = x\})$.

Example. $X(i, j) = i + j$

$$\mathbb{P}(X = 4) = \mathbb{P}(\{(1,3), (3,1), (2,2)\}) = 3 \times \frac{1}{36} = \frac{1}{12}$$

$$\mathbb{P}(X = 3) = \mathbb{P}(\{(1,2), (2,1)\}) = 2 \times \frac{1}{36} = \frac{2}{36} = \frac{1}{18}$$

$$\mathbb{P}(X = 2) = \mathbb{P}(\{(1,1)\}) = 1 \times \frac{1}{36} = \frac{1}{36}$$

Random Variables

Definition. For a RV X, we define the event

$$\{X = x\} = \{\omega \in \Omega \mid X(\omega) = x\}$$

We write $\mathbb{P}(X = x) = \mathbb{P}(\{X = x\})$.

Example. $Z(i, j) = i$

$$\mathbb{P}(Z = 2) = \mathbb{P}(\{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)\}) = \frac{1}{6}$$

Example – Number of Heads

We flip n coins, independently, each heads with probability p

$$\Omega = \{\mathrm{HH}\cdots\mathrm{HH}, \mathrm{HH}\cdots\mathrm{HT}, \mathrm{HH}\cdots\mathrm{TH}, \ldots, \mathrm{TT}\cdots\mathrm{TT}\}$$

$X =$ # of heads

$$\mathbb{P}(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

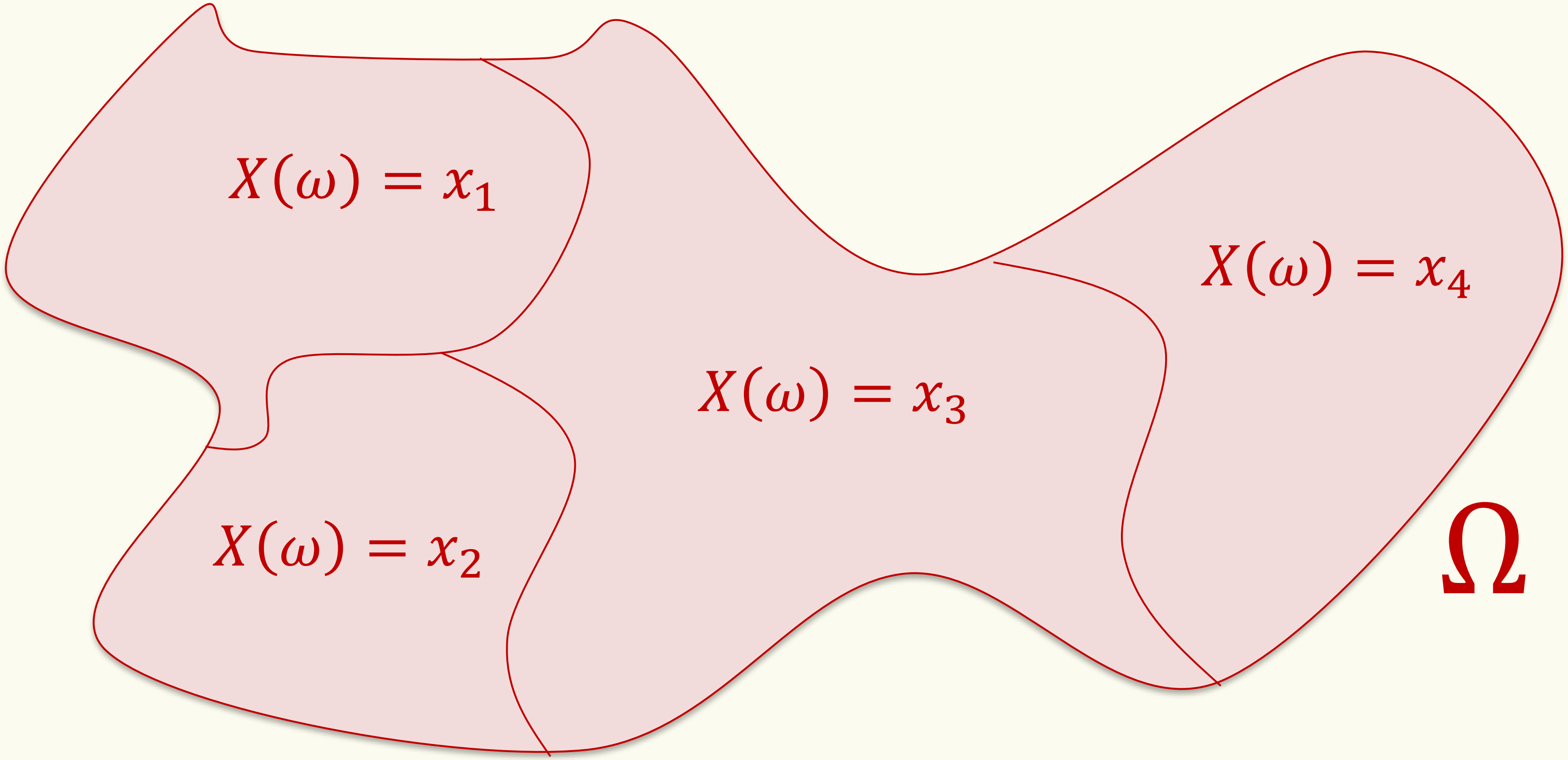
# of sequences with k heads

Prob of sequence w/ k heads

Random Variables and the Probability Space

Random variables partition the sample space.

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



Distribution of Random Variable

Definition. The **range** of a random variable $X: \Omega \to \mathbb{R}$ is

$$X(\Omega) = \{X(\omega) \mid \omega \in \Omega\}$$

i.e., the set of values the random variable can take.

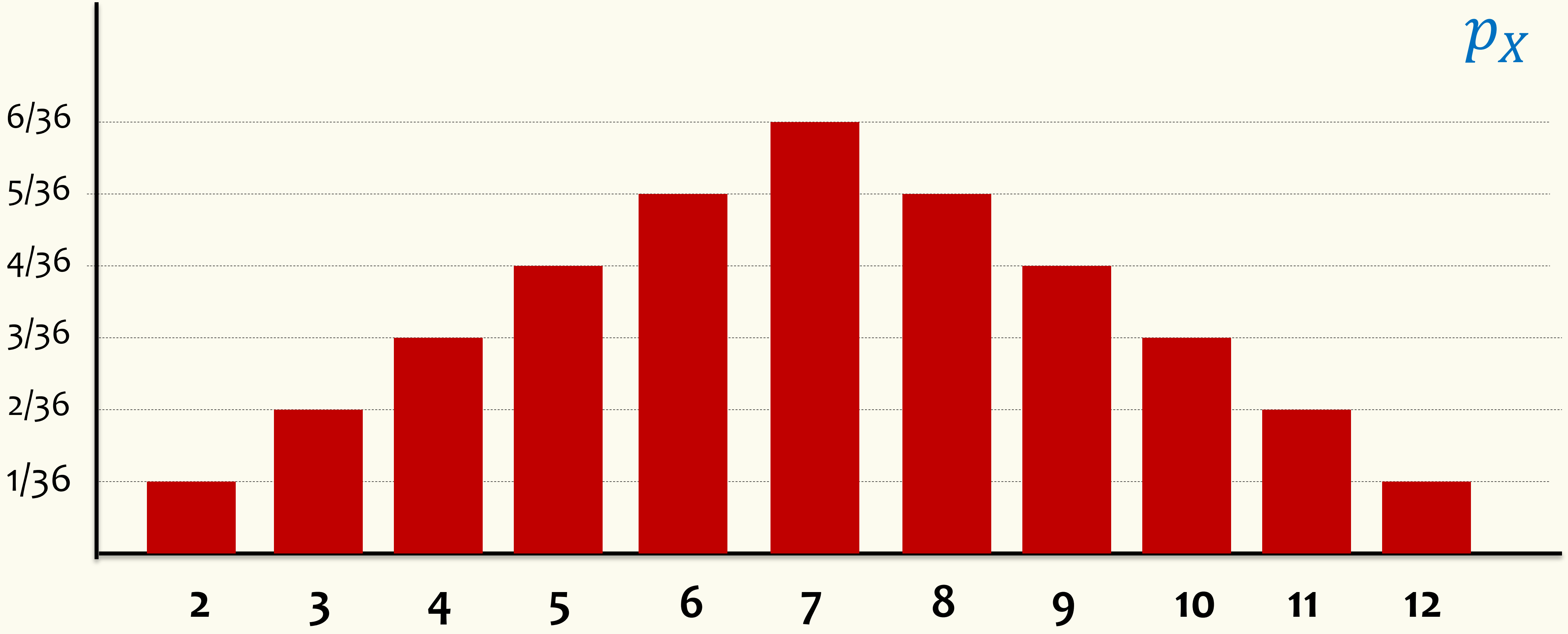
Definition. The **probability mass function (PMF)** of a RV $X: \Omega \to \mathbb{R}$ is the function $p_X: X(\Omega) \to \mathbb{R}$ such that for all $x \in X(\Omega)$:

$$p_X(x) = \mathbb{P}(X = x)$$

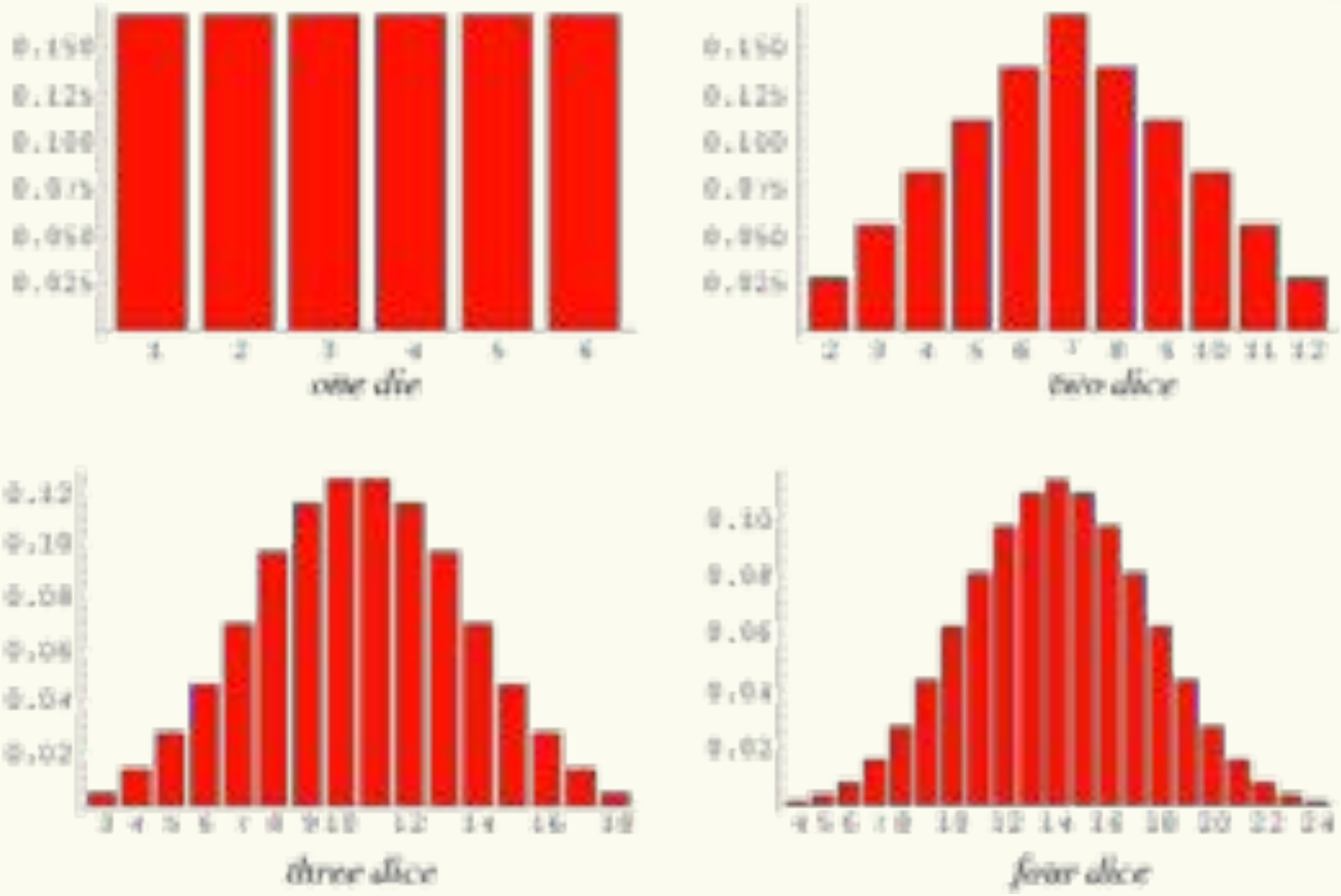
Note: $\sum_x p_X(x) = 1$

Example – Two Dice

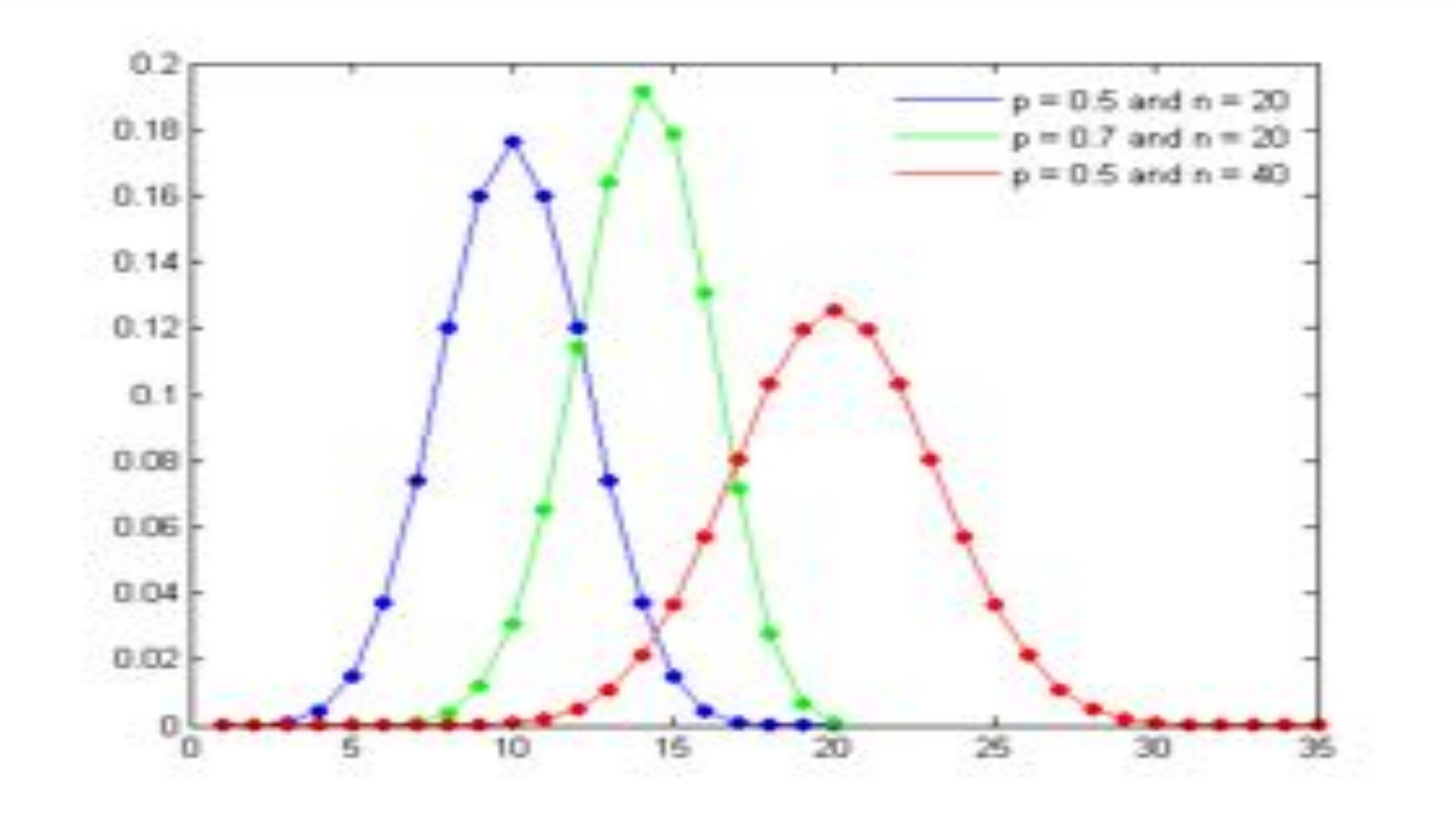
X = sum of two dice throws



Multiple Dice Throw



Example – Number of Heads



$$p_X(k) = \mathbb{P}(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

= **Binomial distribution** with parameters n and p. Denoted $\mathrm{Bin}(n, p)$

Random Variables as Abstraction

- Often, very different probability spaces give rise to random variables with the same distribution.
- We often want to make statements that only depend on the PMF of a random variable, and hence apply to <u>any</u> of these experiments.
- We write $X \sim p$ to say that X is distributed according to p.
  - E.g. $X \sim \mathrm{Bin}(n, p)$