**CS 8850: Advanced Machine Learning** **Fall 2017**

Topic 2: Review of Probability Theory

Instructor: Daniel L. Pimentel-Alarcón 

## 2.1 Why Probability?

Many (if not all) applications of machine learning involve randomness.

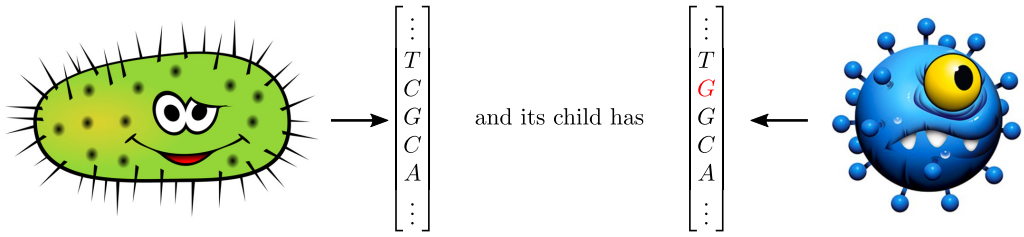
**Example 2.1.** In computer vision we often want to track objects or people. The location of these objects over time is random. For example, look at this frame of a video:

You can predict where some of these people will be in a few seconds (based on their current locations and directions). However, you cannot be certain. What if someone trips? What if the wind moves the leaves of a tree and produces occlusions?) All of this can be modeled probabilistically in a precise mathematical way. For instance, we can say that your predictions will be accurate with a probability p close to 1 (if nothing unusual happens), and inaccurate with probability $1 - \mathrm{p}$ (if something unusual happens). Furthermore, it is harder to predict where these people will be later in time. So we can refine our probability model and say your predictions will be accurate with probability $\mathrm{p}/t$ (where $t$ is the amount of time), and inaccurate with probability $1 - \mathrm{p}/t$.

You can see that *probability theory is nothing but common sense reduced to calculation* — Pierre Laplace, 1812.

**Example 2.2.** All organisms have a DNA sequence (genome), i.e., a very long vector of nucleotides $(A, C, G, T)$. For reasons that nobody knows *yet*, sometimes organisms *mutate*. That is, when organisms

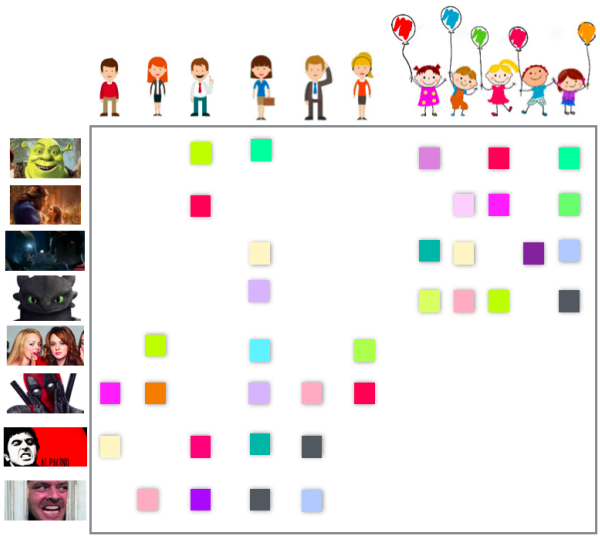
reproduce, some of their nucleotides get changed. For example, it is possible that a parent has the sequence




The location of these mutations are random, but more common in certain regions of the genome (called genes). This can be modeled probabilistically. For instance, we can say that the mutation may appear uniformly at random in each gene, but some genes are more likely to present a mutation than others.

Mutations are more common in smaller organisms, like bacteria and viruses, that reproduce very rapidly. That is why pharmaceuticals, insurance companies, the National Health Service (NHS), the National Security Agency (NSA), and even the Department of Defense (DoD), among many others, are very interested in this problem. The reasons are not as apocalyptic as a zombi virus, but close. For example, some bacteria have become immune to most antibiotics. Similarly, it is harder to produce vaccines for viruses that mutate quickly. In fact these mutations are the reason why we don't have a definitive flu vaccine yet.

**Example 2.3.** Let's now consider Netflix. Some users have rated some movies, some users have rated others. However, nobody has rated all of them. This produces an *incomplete* data matrix like this (colors indicate how much each person liked a movie)



The goal is to predict which users will like which items, in order to make good recommendations. Again, this can be modeled probabilistically. The movies that each user has rated (and hence the samples in this matrix) are somewhat random. For example, adults are more likely to watch (and enjoy) adult

movies, while kids and parents are more likely to watch (and enjoy) children movies. This can be modeled probabilistically.

We could construct similar models for songs, shoes, clothes, restaurants, groceries, etc. So in fact this also applies to Amazon, Pandora, Spotify, Pinterest, Yelp, Apple, etc. If these companies recommend you an item you will like, you are more likely to buy it. You can see why all these companies have a great interest in this problem, and they are paying *a lot* of money to people who work on this.

I hope these few examples help convincing you of the ubiquitousness of randomness in machine learning. Probability theory allows us to model randomness in a precise mathematical way. Furthermore, despite the uncertainty produced by randomness, probability theory allows us to draw *likely* conclusions in a sensible manner, and quantify how certain we are about these conclusions.

## 2.2 The Basics

**Definition 2.1.** There are three elemental concepts in basic probability theory:

$\Omega$ := Sample space = set of all possible outcomes.

$\mathcal{A}$ := Set of all possible events.

$\mathsf{P}$ := Probability measure.

**Example 2.4.** Consider a fair die. Then

$$\begin{aligned}\Omega &= \{1,2,3,4,5,6\}.\\ \mathcal{A} &= \Big\{\{1\},\ldots,\{6\},\{1,2\},\ldots,\{5,6\},\ldots,\{1,2,3,4,5,6\}\Big\}\\ \mathsf{P}\ (\mathrm{k}) &= 1/6 \text{ for every } \mathrm{k} = 1,\ldots,6.\end{aligned}$$

**Definition 2.2** (Probability measure)**.** A mapping $\mathsf{P} : \mathcal{A} \to [0,1]$ is a *probability measure* if it satisfies the next properties:

(i) $\mathsf{P}(A) \geq 0$ for every $A \in \mathcal{A}$.

(ii) $\mathsf{P}(\Omega) = 1$.

(iii) $A \cap B = \emptyset$ for some $A, B \in \mathcal{A}$, then $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B)$.

Condition (iii) implies that $\mathsf{P}(A \cup B) \leq \mathsf{P}(A) + \mathsf{P}(B)$, which is often known as the *union bound*.

## 2.3 Conditional Probability

**Definition 2.3** (Conditional probability)**.** Let $A, B \in \mathcal{A}$. The *conditional probability* that $A$ occurred given $B$ occured is

$$\mathsf{P}(A|B) \;:=\; \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}$$

**Example 2.5.** Continuing with Example 2.4, let $A = \{1, 2\}$, $B = \{2, 3\}$. The probability that $A$ occurs is $\mathsf{P}(A) = 1/3$. However, if you already know that $B$ occurred, then the probability that $A$ also occurred increases to

$$\mathsf{P}(A|B) \;=\; \frac{1/6}{1/3} \;=\; \frac{1}{2}.$$

## 2.4 Independence

**Definition 2.4** (Independent events)**.** Let $A, B \in \mathcal{A}$. We say $A$ and $B$ are *independent* if $\mathsf{P}(A|B) = \mathsf{P}(A)$.

In words, two events are independent if they provide no information of one an other.

**Example 2.6.** Consider two fair dice. Let $A$ be the event that the first die is 1; let $B$ be the event that the second die is 1. Then

$$\mathsf{P}(A|B) \;=\; \frac{\mathsf{P}(A = 1 \cap B = 1)}{\mathsf{P}(B = 1)} = \frac{1/36}{1/6} = \frac{1}{6} = \mathsf{P}(A).$$

Hence the events $A$ and $B$ are independent. This matches our intuition that one die has no influence on the outcome of the other.

## 2.5 Bayes Rule

Given the conditional probability $\mathsf{P}(A|B)$, Bayes rule gives us a formula for the *inverse* probability, $\mathsf{P}(B|A)$.

**Definition 2.5** (Bayes rule)**.** Let $A, B \in \mathcal{A}$. Then

$$\mathsf{P}(B|A) \;=\; \frac{\mathsf{P}(A|B)\mathsf{P}(B)}{\mathsf{P}(A)}$$

Bayes rule plays a crucial role in modern applications.

**Example 2.7.** Geneticists have determined that 90% of the people with disease $B$ have gene $A$ active, i.e., $\mathsf{P}(A|B) = 0.9$. If you sequence your genome and find out that your gene $A$ is active, what is the probability that you develop disease $B$? In other words, what is $\mathsf{P}(B|A)$? At first glance you might think it is very likely that you will develop disease $B$. However, to determine this you need to know $\mathsf{P}(A)$ and $\mathsf{P}(B)$. Of the whole population, if only 5% have disease $B$, while 45% have gene $A$ active, what is $\mathsf{P}(B|A)$? This is a simple application of Bayes rule:

$$\mathsf{P}(B|A) \;=\; \frac{\mathsf{P}(A|B)\mathsf{P}(B)}{\mathsf{P}(A)} \;=\; \frac{(0.9)(0.05)}{0.45} \;=\; 0.1$$

## 2.6 Random Variables

Sometimes the sample space $\Omega$ contains elements that are cumbersome to handle. For example, imagine making a list of all animals, $\Omega = \{\textit{elephant}, \textit{giraffe}, \dots\}$. Other times, $\Omega$ is not explicitly identified. Hence we want to translate from the world of outcomes to a more familiar and measurable space, like $\mathbb{R}$. That is essentially why we use random variables.

**Definition 2.6** (Random variable)**.** A random variable is a mapping $x : \Omega \to \mathbb{R}$.

For example, we could define a mapping $x$ that maps *elephant* $\mapsto 1$, *giraffe* $\mapsto 2$, etc. Since $\mathsf{P}$ specifies a probability for every $A \in \mathcal{A}$, it also induces a probability in terms of $x$. For instance, the event $\{x \leq 0\}$ is equivalent to the event $\{\omega \in \Omega \;:\; x(\omega) \leq 0\}$, and $\mathsf{P}(x \leq 0) = \mathsf{P}(\{\omega \in \Omega \;:\; x(\omega) \leq 0\})$. Continuing with our example, $P(x \leq 2) = \mathsf{P}(\{\textit{elephant}, \textit{giraffe}\})$.

## 2.7 Densities

Intuitively, a probability measure $\mathsf{P}$ is the rule that assigns probability to the events in $\mathcal{A}$. For instance, in Example 2.4, $\mathsf{P}$ assigns an equal probability of $1/6$ to each of the possible outcomes $\mathrm{k} = 1, \dots, 6$. To calculate the probability of an event, all we would have to do is compute

$$\mathsf{P}(x \in A) \;=\; \sum_{\mathrm{k} \in A} \mathsf{P}(x = \mathrm{k}).$$

This was easily done because $x$ was discrete. In this case, $\mathsf{P}$ is called a *mass function*

If $x$ were continuous, the probability that it takes a particular value is zero. So instead of a mass function, we use a *density function* $\mathsf{p}$ that indicates the probability that $x$ falls within certain intervals. Then

$$\mathsf{P}(x \in A) \;=\; \int_A \mathsf{p}(\mathrm{x}) d\mathrm{x}.$$

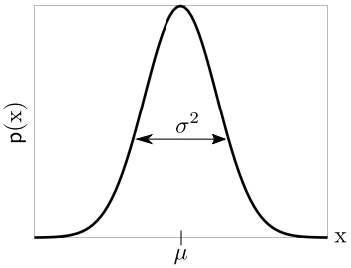



Figure 2.1: Gaussian density function $\mathsf{p}(\mathrm{x})$ with mean $\mu$ and variance $\sigma^2$.

**Example 2.8.** The height of a person can be modeled as a gaussian random variable with mean $\mu = 5'5"$ for females and $\mu = 5'10"$ for males (see Section 2.9 and Figure 2.1). However, the probability that your height is *exactly* $5'5"$ or $5'10"$ is zero. You are more likely to be somewhere in between $(5'3", 5'7")$ or $(5'8", 5'12")$.

## 2.8 Expectation

**Definition 2.7** (Expectation)**.** For a continuous random variable $x$ and an arbitrary function $f(x)$,

$$\mathsf{E}[f(x)] \;:=\; \int f(\mathrm{x})\mathsf{p}(\mathrm{x})d\mathrm{x}.$$

If $x$ is a discrete random variable,

$$\mathsf{E}[f(x)] \;:=\; \sum_{\mathrm{k}} f(\mathrm{x_k})\mathsf{P}(x = \mathrm{x_k}).$$

**Example 2.9.** Special cases of expectations:

- **Probability:** $\mathsf{P}(x \in A) = \mathsf{E}[\mathbb{1}_{x \in A}]$.
- **Mean:** $\mu := \mathsf{E}[x]$.
- **Variance:** $\sigma^2 := \mathsf{E}[(x - \mu)^2]$.

**Definition 2.8** (Conditional expectation)**.** The *conditional expectation* of a continuous random variable $f(x)$ given a random variable $y$ takes the value y is defined as

$$\mathsf{E}[f(x)|y = \mathrm{y}] \;:=\; \int f(\mathrm{x})\mathsf{p}(\mathrm{x}|\mathrm{y})d\mathrm{x},$$

and if $x$ is a discrete random variable,

$$\mathsf{E}[f(x)|y = \mathrm{y}] \;:=\; \sum_{\mathrm{k}} f(\mathrm{x_k})\mathsf{P}(x = \mathrm{x_k}|\mathrm{y}),$$

## 2.9 Common Probability Measures

Tables 2.1 and 2.2 give examples of common probability measures, also known as *distributions*.

## 2.10 Multivariate distributions

In many modern applications it is convenient to arrange random variables in vectors. For example, we might consider a *random vector*

$$\boldsymbol{x} \;=\; \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

containing the information of a person's height, weight and cholesterol level. If the random variables in the vector are independently distributed, then its *joint* distribution is just the product of the univariate distributions of each component. In our example, $\mathsf{p}(\mathbf{x}) = \mathsf{p}(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)$ would simply factor into $\mathsf{p}(\mathrm{x}_1)\mathsf{p}(\mathrm{x}_2)\mathsf{p}(\mathrm{x}_3)$. However, if the random variables in the vector are dependent (as is actually the case with height, weight and cholesterol level), then $\mathsf{p}(\mathbf{x})$ does not factor in this simple way.

Multivariate densities model dependent random variables. The one we will use most in this course is the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{D}}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$, which has the following form:

$$\mathsf{p}(\mathbf{x}) \;=\; \frac{1}{\sqrt{(2\pi)^{\mathrm{D}}|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

| **Discrete** | $\mathsf{P}(\mathrm{x})$ | $\mathsf{E}[x]$ | $\mathsf{var}(x)$ |
|---|---|---|---|
| Bernoulli(p) | $\mathsf{P}(x = 1) = \mathrm{p}$, $\mathsf{P}(x = 0) = 1 - \mathrm{p}$ | p | $\mathrm{p}(1 - \mathrm{p})$ |
| Binomial(n, p) | $\mathsf{P}(x = \mathrm{k}) = \binom{\mathrm{n}}{\mathrm{k}} \mathrm{p}^{\mathrm{k}}(1 - \mathrm{p})^{\mathrm{n}-\mathrm{k}}$, $\mathrm{k} = 0, \ldots, \mathrm{n}$ | np | $\mathrm{np}(1 - \mathrm{p})$ |
| Poisson($\lambda$) | $\mathsf{P}(x = \mathrm{k}) = \frac{\lambda^{\mathrm{k}} e^{-\lambda}}{\mathrm{k!}}$, $\mathrm{k} = 0, 1, \ldots$ | $\lambda$ | $\lambda$ |

Table 2.1: Examples of common probability mass functions.

| **Continuous** | $\mathsf{p}(\mathrm{x})$ | $\mathsf{E}[x]$ | $\mathsf{var}(x)$ |
|---|---|---|---|
| Uniform[a, b] | $\mathsf{p}(\mathrm{x}) = \frac{1}{\mathrm{b}-\mathrm{a}}$, $\mathrm{a} \leq \mathrm{x} \leq \mathrm{b}$ | $\frac{\mathrm{a}+\mathrm{b}}{2}$ | $\frac{(\mathrm{b}-\mathrm{a})^2}{12}$ |
| Exponential($\lambda$) | $\mathsf{p}(\mathrm{x}) = \lambda e^{-\lambda \mathrm{x}}$, $\mathrm{x} \geq 0$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Laplace($\lambda$) | $\mathsf{p}(\mathrm{x}) = \frac{\lambda}{2} e^{-\lambda \lvert \mathrm{x} \rvert}$ | 0 | $\frac{2}{\lambda^2}$ |
| Gaussian or Normal($\mu, \sigma^2$) | $\mathsf{p}(\mathrm{x}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu}{\sigma}\right)^2}$ | $\mu$ | $\sigma^2$ |
| Gamma($\alpha, \beta$) | $\mathsf{p}(\mathrm{x}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{x}^{\alpha-1} e^{-\beta \mathrm{x}}$ <br> $\Gamma(\alpha) := \int_0^\infty \mathrm{x}^{\alpha-1} e^{-\mathrm{x}} d\mathrm{x}$ | $\frac{\alpha}{\beta}$ | $\frac{\alpha}{\beta^2}$ |
| Beta($\alpha, \beta$) | $\mathsf{p}(\mathrm{x}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mathrm{x}^{\alpha-1}(1-\mathrm{x})^{\beta-1}$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ |
| $\chi^2(\mathrm{k})$ | Gamma($\mathrm{k}/2, 1/2$) | | |
| $F(\alpha, \beta)$ | $\mathsf{p}(\mathrm{x}) = \frac{\sqrt{\frac{(\alpha \mathrm{x})^\alpha \beta^\beta}{(\alpha \mathrm{x}+\beta)^{\alpha+\beta}}}}{\mathrm{x}\mathcal{B}(\alpha/2, \beta/2)}$ <br> $\mathcal{B}(\alpha, \beta) := \int_0^1 \mathrm{x}^{\alpha-1}(1-\mathrm{x})^{\beta-1} d\mathrm{x}$ | $\frac{\beta}{\beta-2}$, $\beta > 2$ | $\frac{2\beta^2(\alpha+\beta-2)}{\alpha(\beta-2)^2(\beta-4)}$, $\beta > 4$ |

Table 2.2: Examples of common probability density functions.

## 2.11 Sums of Independent Random Variables

In many applications we want to know the distribution of the sum of independent random variables. Table 2.3 gives a few examples.

**Example 2.10.** Suppose there is an epidemic in a city with N habitants. The $\mathrm{i}^{\mathrm{th}}$ person will independently contract the disease with probability p. We can model this as $x_{\mathrm{i}} \overset{iid}{\sim} \text{Bernoulli(p)}$, $\mathrm{i} = 1, \ldots, \mathrm{N}$. Let $m = \sum_{\mathrm{i}=1}^{\mathrm{N}} x_{\mathrm{i}}$ be the number of people that get infected. The Center for Disease Control wants to determine $\mathsf{P}(m > \mathrm{m})$. This raises the question: what is the distribution of $m$? Notice that $m$ is the sum of i.i.d. Bernoulli random variables. However, $m$ is clearly not Bernoulli. To begin with, $m$ can take values in $\{0, \ldots, \mathrm{N}\}$, while a Bernoulli random variable can only take values in $\{0, 1\}$. So the question is: what is the distribution of a sum of N i.i.d. Bernoulli(p) random variables?

## 2.12 Other Common Functions of Random Variables

In addition to sums, other common functions of random variables include

- **Linear multiplication:** For a constant matrix $\mathbf{A} \in \mathbb{R}^{\mathrm{M}\times\mathrm{D}}$ and a random vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$, $\mathsf{cov}(\mathbf{A}\boldsymbol{x}) = \mathbf{A}\mathsf{cov}(\boldsymbol{x})\mathbf{A}^{\mathsf{T}}$.
- **Squared gaussians:** If $x \sim \mathcal{N}(0, 1)$, then $x^2 \sim \chi^2$.
- **Ratio of $\chi^2$'s:** If $x \sim \chi^2(\mathrm{k})$ is independent of $y \sim \chi^2(\ell)$, then $\dfrac{\ell x}{\mathrm{k} y} \sim F(\mathrm{k}, \ell)$.

| $x_\mathrm{i}$ | $\sum_{\mathrm{i}=1}^{\mathrm{N}} x_\mathrm{i}$ |
|---|---|
| Bernoulli(p) | **Exercise** |
| Binomial($\mathrm{n_i}$, p) | $\mathrm{Binomial}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{n_i}, \mathrm{p}\right)$ |
| Poisson($\lambda_\mathrm{i}$) | $\mathrm{Poisson}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \lambda_\mathrm{i}\right)$ |
| exp($\lambda$) | Gamma(N, $\lambda$) |
| Gamma($\alpha_\mathrm{i}, \beta$) | $\mathrm{Gamma}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_\mathrm{i}, \beta\right)$ |
| $\mathcal{N}(\mu_i, \sigma_i^2)$ | $\mathcal{N}\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mu_\mathrm{i}, \sum_{\mathrm{i}=1}^{\mathrm{N}} \sigma_\mathrm{i}^2\right)$ |
| $\chi^2(\mathrm{k_i})$ | $\chi^2\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{k_i}\right)$ |

Table 2.3: Examples of distributions of sums of **independent** random variables.