Probability Notes

0 Vocabulary

The purpose of this lecture to become familiar with these terms:

- Probability
- Random variable
- Probability distribution
- Conditional probability
- Binary classifier
- Classification threshold
- Receiver operating characteristic (ROC)
- Area under ROC (AUROC)

1 Probability, random variables, and distributions

1.1 Modeling a fair coin

Probability is a mathematical framework for describing uncertain events, and understanding their properties. For example: when flipping a coin, it is hard to predict whether a coin will land on 'heads' or 'tails'. To respect this fact, for a fair coin, we say:

- There is a 1/2 **probability** of the coin landing on heads.
- There is a 1/2 **probability** of the coin landing on tails.

Let's translate this description into the mathematical formalism of probability.

1. Name the outcome of the coin flip X. Call X a **random variable**.
2. Determine all the possible values that X can have. In this case: *heads* or *tails*.
3. Assign a probability to each possible value. In this case, we can write

$$\begin{aligned}\Pr(X = \text{heads}) &= 1/2\\ \Pr(X = \text{tails}) &= 1/2\end{aligned}$$

where Pr(*something something*) is read as 'the probability of *something something*':

$$\Pr(X = \text{heads}) \qquad \text{is read as} \qquad \text{the probability that } X \text{ equals heads}$$

This is known as the **probability distribution** of X. Probabilities can be any value between 0 and 1, but the sum of the probabilities of all possible values of the random variable must be 1.

That's it!

This notation is very general. You may be surprised by *heads* and *tails* appearing as words in the formulas above, *but the statement above is rigorous math.* In general, random variables can take many kinds of value, such as: heads, tails, yes, no, apple, cat, dog, 1, 2, 3, 4, 5, $\ldots$, 0.1, 0.01, 0.001, π, e, α, β, and tuples like $(\pi, \text{yes}, 0)$. It just depends on the problem being analyzed.

1.2 Example: the binomial distribution

Let's see a slightly more complicated example of a distribution. Consider flipping a fair coin 15 times, and counting how many heads you see across all 15 flips. Since each flip is random, the total number of heads is also random. Let's call it Y:

$$Y \quad \equiv \quad \text{the number of heads seen in 15 flips of a fair coin}$$

Y can take any integer value between 0 and 15.

- If $Y = 0$, then all flips came up tails.
- If $Y = 15$, then all flips came up heads.

It turns out that the **probability distribution** of Y is:

$$\begin{aligned}
\Pr(Y = 0) \quad & = \frac{1}{2^{15}} && \approx 0.00003 \\
\Pr(Y = 1) \quad & = \frac{1}{2^{15}}\binom{15}{1} && \approx 0.0004 \\
\Pr(Y = 2) \quad & = \frac{1}{2^{15}}\binom{15}{2} && \approx 0.003 \\
& \vdots \\
\Pr(Y = 14) \quad & = \frac{1}{2^{15}}\binom{15}{14} && \approx 0.0004 \\
\Pr(Y = 15) \quad & = \frac{1}{2^{15}} && \approx 0.00003
\end{aligned}$$

Or:

$$\Pr(Y = k) = \frac{1}{2^{15}} \cdot \binom{15}{k} \qquad \text{for all } k = 0, 1, 2, \ldots, 15$$

(This is a special case of the binomial distribution. Don't remember it! Just notice that sometimes it is easier to specify a probability distribution with a formula than to write down every probability individually.)

Suppose we only care about whether Y is 8 or more. We can calculate the probability of that event as

$$\begin{aligned}
\Pr(Y \geq 8) \quad & = \quad \Pr(Y = 8 \text{ or } Y = 9 \text{ or } \ldots \text{ or } Y = 15) \\
& = \quad \Pr(Y = 8) + \Pr(Y = 9) + \cdots + \Pr(Y = 15) \\
& \big(= 1/2 \big)
\end{aligned}$$

In addition, since probabilities always add to 1, we have

$$\Pr(Y > 0) = 1 - \Pr(Y = 0)$$

and

$$\Pr(Y > 1) = 1 - \Pr(Y = 0) - \Pr(Y = 1)$$

and so on.

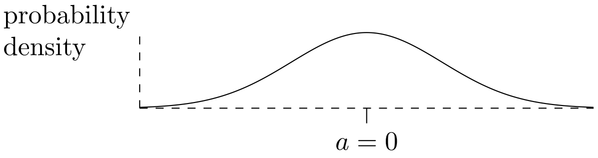
> **Exercise 1**
>
> Let R be a random variable modeling the value of a dice roll. The die is a similar normal six-sided die, but someone painted a '5' to replace the '2'.
>
> 1. What is the probability distribution of R?
> 2. What is the probability that R is even?

1.3 Continuous random variables

Random variables can also take on continuous values. This is especially useful for modeling measurements, such as weight, voltage, rate, etc., as random variables.

Most common distributions for continuous random variables X can be described with probability density functions $f(a)$ where $\Pr(X \text{ is near } a)$ is proportional to $f(a)$. The mathematical details are trickier but the intuition is the same as the discrete case.

The most common continuous distribution is the 'standard normal distribution':



which has a probability density function

$$f(a) = \frac{1}{\sqrt{2\pi}} e^{-a^2/2}$$

From the probability density plot, we can see that the standard normal distribution is symmetric and concentrated around $a = 0$.

2 Conditional probability

2.1 A drug trial

Often we are interested in two (or more) random variables at once, and particularly in how they interact with each other. One way of quantifying this interaction is through **conditional probability.**

Suppose we are testing a drug which is meant to manage symptoms of some disease. Every time a new participant joins the study, we flip a weighted coin:

- If the coin lands on heads, we administer a placebo (P)
- If the coin lands on tails, we administer the drug (D).

One month later, we follow up with the participant to assess whether their symptoms got better (S) or worse (F). Then, over many participants, we record the administered substance and the result, getting a table like:

Participant	Substance	Result
1	D	S
2	D	F
3	P	F
4	D	S
5	P	F
6	P	S
⋮	⋮	⋮

For each participant, there are only four values for what we write down:

D, S D, F P, S P, F

So we can tally up the number of times that each of these possibilities occurs:

Substance	Result	Count
D	S	80
D	F	30
P	S	20
P	F	70

(There are 200 participants in the study.) Now, let's model this result using random variables.

- Let A be a random variable corresponding to the substance we administered: D (drug) or P (placebo).
- Let R be a random variable corresponding to the medical result: S (success) or F (failure).

> **Exercise 2**
>
> *Given this table*, what is the probability distribution of A? What is the probability distribution of R?

What we really want to know about the drug is: does it work better than the placebo? To answer this question, we need to compare the conditional probabilities:

(1) The probability of success given we administered the placebo
(2) The probability of success given we administered the drug

In the random variable notation, these are written:

(1) $\Pr(R = \mathrm{S}|A = \mathrm{P})$
(2) $\Pr(R = \mathrm{S}|A = \mathrm{D})$

(just read the '|' as 'given'.)

The **conditional probability** is defined as:

$$\Pr(R = x|A = y) = \frac{\Pr(R = x \text{ and } A = y)}{\Pr(A = y)}$$

so therefore

$$\Pr(R = S|A = P) = \frac{20/200}{(20 + 70)/200} = 2/9$$
$$\Pr(R = S|A = D) = \frac{80/200}{(80 + 30)/200} = 8/11$$

so in this case, the drug is better than the placebo.

> **Exercise 3**
>
> Calculate
>
> 1. $\Pr(A = P|R = S)$
> 2. $\Pr(A = D|R = S)$
>
> Notice that, generally:
>
> $$\Pr(X = x|Y = y) \neq \Pr(Y = y|X = x)$$

2.2 Independence

Random variables X and Y are independent if and only if for all values x and y,

$$\Pr(X = x|Y = y) = \Pr(X = x) \quad \text{and} \quad \Pr(Y = y|X = x) = \Pr(Y = y)$$

> **Exercise 4**
>
> Suppose X and Y are independent. What is $\Pr(X = x \text{ and } Y = y)$? (Start from the def. above!)

2.3 Bayes' rule

We have

$$\Pr(X = a | Y = b) = \frac{\Pr(X = a \text{ and } Y = b)}{\Pr(Y = b)}$$

and

$$\Pr(Y = b | X = a) = \frac{\Pr(X = a \text{ and } Y = b)}{\Pr(X = a)}$$
$$\implies \Pr(X = a \text{ and } Y = b) = \Pr(Y = b | X = a) \Pr(X = a)$$

Substituting this into the first equation, we get:

$$\Pr(X = a | Y = b) = \frac{\Pr(Y = b | X = a) \Pr(X = a)}{\Pr(Y = b)}$$

This is known as **Bayes' rule**.

> **Exercise 5**
>
> Suppose you are a doctor.
>
> - One day, a patient comes in displaying symptom LPS.
> - You know that 10% of patients displaying LPS have lupus, so you administer a lupus test.
> - It is a very accurate test: for 90% of patients with lupus it comes out positive, and for 90% of healthy patients it comes out negative.
>
> Formally:
>
> - Let L be a random variable representing whether the patient has lupus: we know
>
> $$\Pr(L = \text{lupus}) = 0.1 \quad \text{and} \quad \Pr(L = \text{healthy}) = 0.9$$
>
> - Let T be a random variable representing the test result: we know
>
> $$\Pr(T = \text{positive} | L = \text{lupus}) = \Pr(T = \text{negative} | L = \text{healthy}) = 0.9$$
>
> *Part 1.* Calculate $\Pr(T = \text{positive})$. Hint: use
>
> $$\Pr(T = \text{positive}) = \Pr(T = \text{positive and } L = \text{lupus}) + \Pr(T = \text{positive and } L = \text{healthy})$$
>
> *Part 2.* The test comes out positive! Given this test result, what is the probability that the patient has lupus?

3 Binary classifiers and ROC analysis

3.1 Binary classifiers

In the last exercise, the test is a **binary classifier** because it classifies patients into two groups (healthy/lupus). Binary classifiers are extremely common. The general setup is something like:

- We have an object which we believe falls into one of two classes, such as 'yes' and 'no'. We can model this class as a random variable C.
- We measure some data, which we model as a random variable D. Depending on the specific case, the values taken by D can really be anything! (E.g.: 1000-dimensional vectors).
- Our **binary classifier** is a function f so that $f(D)$ is yes or no, depending on the value of D. The output of the binary classifier is another random variable: $X = f(D)$.

A binary classifier is accurate if $\Pr(X = C)$ is close to 1. (For a perfect classifier, $\Pr(X = C) = 1$.)

Comparing X and C, there are only four things that can happen:

	$X = \text{yes}$	$X = \text{no}$
$C = \text{yes}$	true positive	false negative
$C = \text{no}$	false positive	true negative

People have named various probabilities relating to X and C:

name	expression
accuracy	$\Pr(X = C)$
true positive rate (TPR)	$\Pr(X = \text{yes} \mid C = \text{yes})$
false positive rate (FPR)	$\Pr(X = \text{yes} \mid C = \text{no})$
false negative rate (FNR)	$\Pr(X = \text{no} \mid C = \text{yes}) = 1 - \text{TPR}$
true negative rate (TNR)	$\Pr(X = \text{no} \mid C = \text{no}) = 1 - \text{FPR}$

...and many more. See `https://en.wikipedia.org/wiki/Receiver_operating_characteristic` for a nice table.

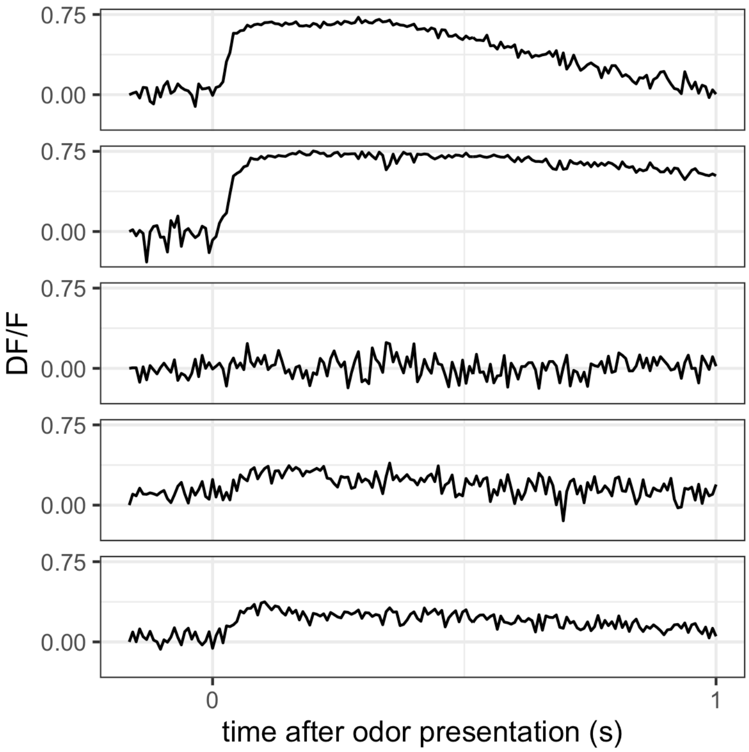
Note: accuracy depends on the prevalences of the ground truth classes, i.e. the distribution of C. On the other hand, measures like TPR/FPR/FNR/TNR do not.

3.2 Threshold classifiers

Almost all classifiers work by calculating a (continuous valued) *score* from D, and then thresholding on that score to decide X. Explicitly, for a given threshold α:

$$X = \begin{cases} \text{yes} & \text{score}(D) > \alpha \\ \text{no} & \text{score}(D) \leq \alpha \end{cases}$$

For example: suppose you are doing nuclear calcium fluorescence imaging of odor-encoding neurons in fruit flies. You record 1000 neurons at a time. These neurons are normally silent, but respond sparsely to odors. After a specific odor is presented to the fly, neurons that are tuned to that odor will fire, which is reflected by an increase in the fluorescence signal over the next second. Here's some example signals:



(Each trace is one neuron.)

We want to make a classifier that decides whether a neuron responds to the odor. To do this, we can:

1. Define a *score* function, which converts each trace above into a single number.
2. Choose a **threshold** score α which separates responsive neurons from non-responsive neurons.

For simplicity, let's choose:

$$\text{score(trace)} = \text{the max DF/F achieved by the trace}$$

This gives us a score table like:

trace	score
1	0.72
2	0.75
3	0.24
4	0.39
5	0.38

Consider thresholds $\alpha = 0.3$ and $\alpha = 0.5$. Then the classifier X would give:

trace	score	X for $\alpha = 0.3$	X for $\alpha = 0.5$
1	0.72	yes	yes
2	0.75	yes	yes
3	0.24	no	no
4	0.39	yes	no
5	0.38	yes	no

3.3 Receiver operator characteristic (ROC)

How should we choose the classifier threshold? It helps to have some ground truth data. Let's add some labels to our table from before (and a few more datapoints):

trace	score	$C \equiv$ neuron fired?
1	0.72	yes
2	0.75	yes
3	0.24	no
4	0.39	no
5	0.38	yes
6	0.55	yes
7	0.30	yes
8	0.42	no
9	0.28	yes
10	0.61	no

One way to choose the threshold is to pick a bank of *candidate thresholds*, such as

$$0.1 \qquad 0.2 \qquad 0.3 \qquad 0.4 \qquad 0.5 \qquad 0.6 \qquad 0.7 \qquad 0.8 \qquad 0.9 \qquad 1.0$$

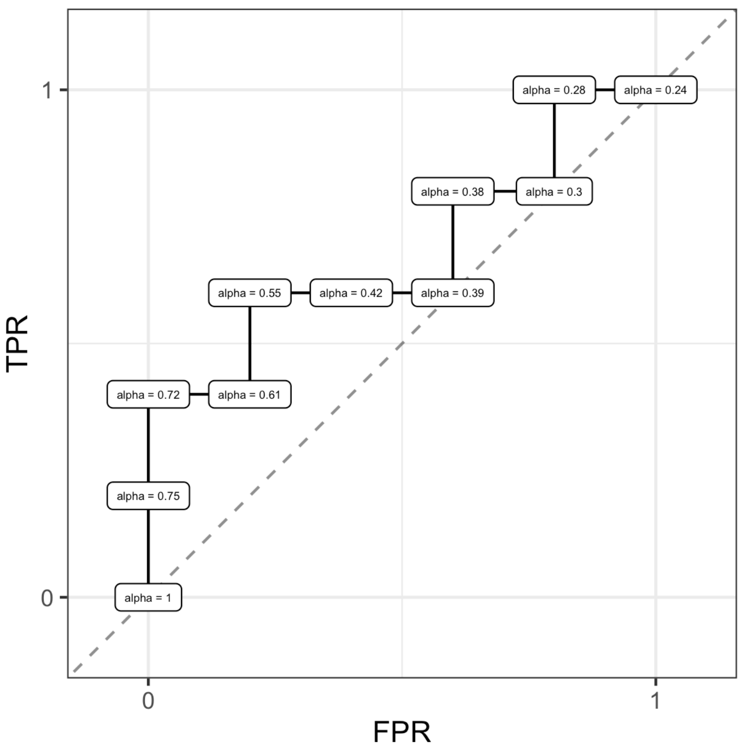
and then calculate the TPR (true positive rate) and FPR (false positive rate) of the classifier assuming each possible value for the threshold.

> **Exercise 6**
>
> In this example, calculate the TPR and FPR for $\alpha = 0.35$.

By the way that the classifier works:

- Increasing α will always *decrease* both the TPR and the FPR. (Or leave them the same.)
- Decreasing α will always *increase* both the TPR and the FPR. (Or leave them the same.)

Then, plot a curve of the TPR vs the FPR. In this case, it looks like:



This plot is called an **receiver operator characteristic (ROC)**. Each possible value of α corresponds to a point on the curve.

> **Exercise 7**
>
> 1. What is the TPR for $\alpha = \infty$? What about the FPR?
> 2. What about for $\alpha = -\infty$?

If score(D) is independent of C, and thus not useful for predicting it, then we will get TPR = FPR for all thresholds. This corresponds to the dashed line $y = x$ in the plot above.

Classifiers that are better than chance have ROC curves that lie above the $y = x$ line. (Classifiers that are worse than chance lie below it.)

You can inspect this curve to choose a classifier with a desirable balance of TPR and FPR. This depends a lot on the application!

- If positives are rare and you really don't want to miss them: choose a threshold with a high TPR, even (potentially) at the expense of a high FPR.
- If you just want to pick out a few examples that are definitely positive (e.g., when choosing raw data to show in a paper): choose a threshold with a low FPR, even (potentially) at the expense of a low TPR.

3.4 AUROC

Finally: sometimes we just want a number to quantify whether our score function is able to separate positives from negatives, and we don't really care about choosing a threshold or using the resulting classifier. In this case, people often calculate the **area under the ROC (AUROC)** as a way to 'grade' the score function. The AUROC is always between 0 and 1, and:

- AUROC > 0.5: the score function is better than chance at distinguishing positive/negative examples.
- AUROC $= 0.5$: the score function is no better than chance.
- AUROC < 0.5: worse than chance.

There is also a precise interpretation of the AUROC in terms of probability: for two independent samples, with independent scores and classes,

$$\text{AUROC} = \Pr(\text{score}_1 > \text{score}_2 \mid C_1 = \text{yes and } C_2 = \text{no})$$

In words: the AUROC is the probability that a random positive example has a higher score than a random negative example.