Week 2 - 1/2: Decision Theory

Motivation: How to make decisions?

Why do we care about probabilities in the first place?
Answer: They help us make decisions.

Pascal, 1670: When faced with a choice of actions, you should:

1. Determine the value (goodness) of all possible outcomes. $V(o) \quad \forall o$
2. Dind the probability of each outcome under each action. $p(o|a) \quad \forall o \forall a$
3. Choose the action with the highest expected value: $\text{argmax}_a \mathbb{E}_{\mathbb{p}(o|a)}[V(o)]$

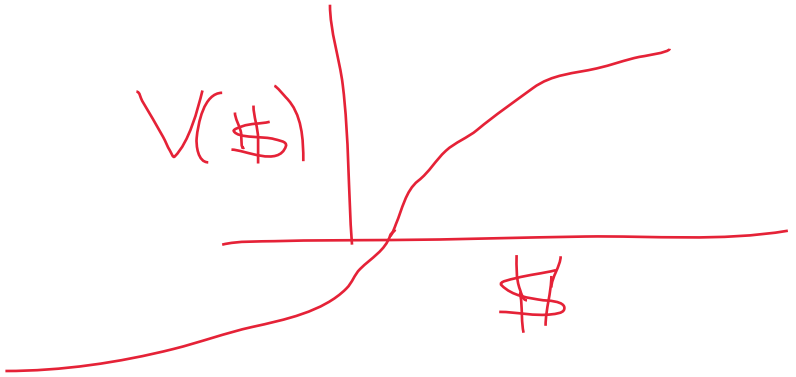
$= \underset{a}{\text{argmax}} \sum_o p(o|a) V(o)$

Example:

I think I might have a bacterial infection. Should I

1. Do nothing,
2. Take penicillin,
3. Take more effective but toxic last-line antibiotics?

Outcome	value
No infection, no toxins.	0 days sick
Infection, no toxins.	5 days sick
No infection, but toxins.	1 day sick
Infection and toxins.	8 days sick



P(outcome given action):	Nothing	Penicillin	Last-line
No infection, no toxins.	0.1	0.9	0
Infection, no toxins.	0.9	0.1	0
No infection, toxins.	0	0	0.99
Infection and toxins.	0	0	0.01
Total	1	1	1

1. Do nothing: 0.9 * 5 = 4.5 expected days sick.
2. Take penicillin 0.1 * 5 = 0.5 expected days sick.
3. Take toxic last-line antibiotics: 0.99 *1 + 0.01* 8 = 1.07 expected days sick.

Objections:

A few common objections to this framework:

Objection 1: I don't care about the average outcome, somes outcomes are smply unacceptable.

Answer: You can't guarantee anything, you can only make probabilities small.

Objection 2: You can't compare the pain of being sick to the cost of medicine in dollars.

Answer: We have to.

The World Health Organization estimated the relative quality people assigned to their own lives under different disabilities:

Condition	Life discount factor
Dementia	0.666
Blindness	0.594
Schizophrenia	0.528
AIDS, not on ART	0.505
Burns 20%-60% of body	0.441
Fractured femur	0.372
Moderate depression episode	0.350
Amputation of foot	0.300
Deafness	0.229
Infertility	0.180
Amputation of finger	0.102

Condition	Life discount factor
Lower back pain	0.061

Objection 3: It's computationally expensive to compute conditional probabilities and expectations over all possible outcomes.

$p(o|a) = \int p(o|a,s)p(s)ds$

state of a situation

Answer: Agreed!

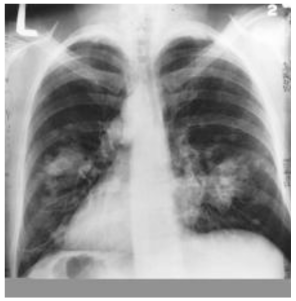
Where did $P(\text{ outcome } | \text{ action })$ come from?
We can always make the model more detailed to include more information.

Probabilities let us make informed decision to make our lives better.

Example: Cancer screening.

x is the set of pixel intensities in the x-ray image.

c represents the presence of cancer, class $\mathcal{C}_1$, or absence of cancer, class $\mathcal{C}_2$.



Our belief after seeing the x-ray is given by:

$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x)} \quad \text{Bayes' rule.}$$

If our goal were simply to make as few misclassifications as possible, we could minimize the expected number of mistakes we'll make.

But a false positve usually has a much different cost than a false negative.

Expected loss

- We want a **loss function** to measure of loss of the decisions under each state of the world.
- Suppose that for some x, the true class is $\mathcal{C}_k$, but we assign x to class $\mathcal{C}_j$. Then we define the loss incurred to be $L(k, j)$.

Here's an example:

Truth \ Decision	cancer	normal
cancer	0	1000
normal	1	0

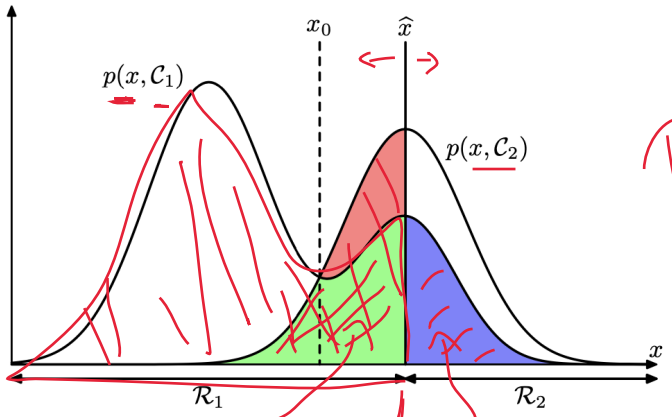
Incorrectly classify as healthy

Incorrectly classify as cancer

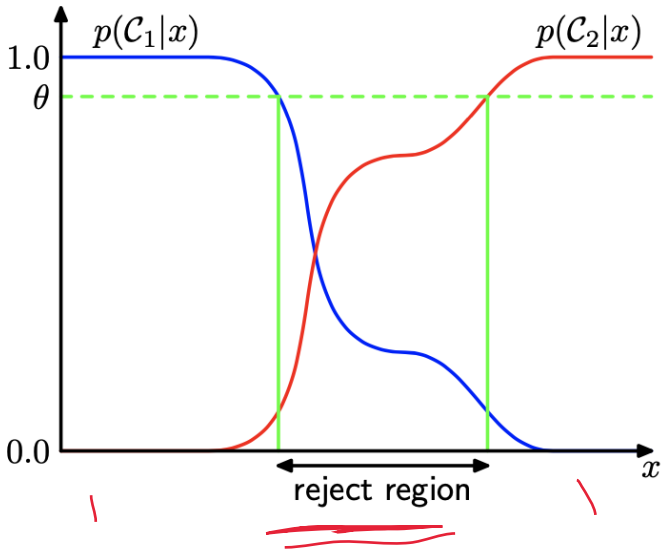
Thus the expected loss is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L(k, j)\, p(x, \mathcal{C}_k) dx$$

Goal is to choose regions $\mathcal{R}_j$ as to minimize expected loss.



$$p(C_1|x) = \frac{p(C_1|x)}{p(C_1|x) + p(C_2|x)}$$



Here's roughly what the conditional distribution looks like for this example:

Sometimes you don't have to make a decision and can take a third option.

In the above figure, the blue region corresponds to $L(1,2)$: the sample comes from class $\mathcal{C}_1$ but we classified as $\mathcal{C}_2$.

Therefore, we want to minimize

$$\begin{aligned}\mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L(k,j)\, p(x,\mathcal{C}_k) dx \\ &= \sum_j \int_{\mathcal{R}_j} \sum_k L(k,j)\, p(x,\mathcal{C}_k) dx.\end{aligned}$$

Define $g_j(x) = \sum_k L(k,j)\, p(x,\mathcal{C}_k)$ and notice that $g_j(x) \geq 0$. Then, the expected loss is equal to

$$\mathbb{E}[L] = \sum_j \int_{\mathcal{R}_j} g_j(x) dx$$

is equivalent to choosing

$$\mathcal{R}_j = \{x \ : \ g_j(x) < g_k(x) \ \text{ for all } \ k \neq j\}.$$

We can also you the product rule $p(x,\mathcal{C}_1) = p(\mathcal{C}_1|x)p(x)$ and reduce the problem to:

Find regions $\mathcal{R}_j$ such that the following is minimized:

$$\sum_k L(k,j)\, p(\mathcal{C}_k|x) \quad \forall j$$

That is

$$\mathcal{R}_j = \{x \ : \sum_k L(k,j)\, p(\mathcal{C}_k|x) < \sum_k L(k,i)\, p(\mathcal{C}_k|x) \ \text{ for all } \ i \neq j\}.$$

R_j = region where action j is best

Tutorial: Loss functions for regression

So far, we have discussed decision theory in the context of classification. We will briefly extend this for the regression setup as well. Now we consider an input/target setup (x, t) where the target (output) is continuous $t \in \mathbb{R}$, and their joint density is given by $p(x, t)$.

Instead of trying to find decision regions, this time we try to find a regression function $y(x) \approx t$ which maps inputs to the outputs. We choose a loss function L between the regression function $y(x)$ and the target t to assess the quality of our estimate, i.e., $L = L(y(x), t)$.

The expected loss in this case,

$$\mathbb{E}[L] = \int \int L(y(x), t) p(x, t) dx dt.$$

> What is the best regression function $y(x)$ that minimizes the expected loss?

Let's choose the loss function as the *squared error loss* to simplify things a bit, i.e. $L(y(x), t) = (y(x) - t)^2$.

$$\begin{aligned}
\mathbb{E}[L] &= \int \int (y(x) - t)^2 p(x, t) dx dt \\
&= \int \int (y(x) - \mathbb{E}[t|x] + \mathbb{E}[t|x] - t)^2 p(x, t) dx dt \\
&= \int \int (y(x) - \mathbb{E}[t|x])^2 p(x, t) dx dt + \int \int (\mathbb{E}[t|x] - t)^2 p(x, t) dx dt
\end{aligned}$$

The last step follows since

$$\begin{aligned}
&\int \int (y(x) - \mathbb{E}[t|x])(\mathbb{E}[t|x] - t) p(x, t) dx dt \\
&= \int \int (y(x) - \mathbb{E}[t|x])(\mathbb{E}[t|x] - t) p(t|x) p(x) dx dt \\
&= \int (y(x) - \mathbb{E}[t|x]) \Big\{ \int (\mathbb{E}[t|x] - t) p(t|x) dt \Big\} p(x) dx = 0
\end{aligned}$$

since the term in the braces is

$$\int (\mathbb{E}[t|x] - t)p(t|x)dt = \mathbb{E}[t|x] - \mathbb{E}[t|x] = 0.$$

We showed that the expected loss is given by the sum of two **non-negative** terms

$$\mathbb{E}[L] = \int\int (y(x) - \mathbb{E}[t|x])^2 p(x,t)dxdt + \int\int (\mathbb{E}[t|x] - t)^2 p(x,t)dxdt.$$

The second term does not depend on $y(x)$ thus choosing the best regression function $y(x)$ is equivalent to minimizing the first term on the right hand side. Thus, we get

$$y(x) = \mathbb{E}[t|x].$$

The second term is the expectation of the conditional variance of $t|x$. It represents the intrinsic variability of the target data and can be regarded as noise.

$(y(x) - t)^2 = \log \mathcal{N}(t|y(x), \sigma^2) + c$

MSE = likelihood under Gaussian noise

Summary

- Depending on the application, one needs to choose an appropriate loss function.
- Loss function can significantly change the optimal decision rule.
- In case of regression, one can find the optimal map between x and t if one knows the conditional distribution $t|x$. The optimal map under a squared loss is the conditional expectation $\mathbb{E}[t|x]$.