

Relevant literature for the course:

- The ICNC book by Gal Chechik, Lidror Troyanski and Naftali Tishby
- The Elements of Statistical Learning by T. Hastie, R. Tibshirani and J. H. Friedman. Online at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
- Probability and Statistics, Maurice DeGroot
- Pattern Recognition and Machine Learning, Christopher Bishop
- Pattern classification, Richard O. Duda, Peter E. Hart and David Stork

1 Statistical Decision Theory

We begin with a probabilistic framework for making decisions in the world. For example, you wake up in the morning and see it is a bit cloudy. Now, in some cases this might indicate it will rain but not always. Now assume we can either wear boots, take an umbrella or do none of these. How should we act.

The situation is thus that we have some observation in the world (it's cloudy) and there is something we don't know (will it rain). Our course of action depends on what will actually happen.

SDT lets us think formally about such decisions. It describes our knowledge about the worlds via the following objects:

- **World States.** The set of possible world states Ω . For example $\Omega = \{\text{it will rain, it won't rain}\}$. We assume that these span all the possible states of the world (i.e., in each possible world exactly one of these is the state).
- **Observations.** Some property of the world that we get to observe X . Could have continuous values (e.g., $X = \mathbb{R}$), multidimensional (e.g., $X = \mathbb{R}^n$) or discrete (e.g., $X = \{\text{it is cloudy, it is sunny, there's a storm}\}$).
- **Possible actions.** The set of actions we can take A . For example $A = \{\text{take umbrella, take coat, stay at home}\}$.
- **Risk function** (or loss function). Whether we are pleased or not with our action depends on what the real state of the world is (which we do not know before we carry out the action). If it rains and we didn't take an umbrella or wore a coat we will pay a lot (we'll be wet). If we took a coat instead we'd pay somewhat less. If we took an umbrella we'd be happy so we won't pay anything. Formally we define a function $\lambda(\alpha|\omega)$ which gives the price we pay for action α if the state of the world is ω . In the discrete case this is a matrix.

So far we said nothing about probabilities. The first two items are actually random variables and they have corresponding distributions.

- **A probabilistic model of the world.** Assume we know the distributions $p(\omega)$ and $p(x|\omega)$. The distribution $p(\omega)$ is called the prior over Ω . In some cases it is natural to assume we know it while in others (e.g., when the world state is the validity of some scientific theory) it might not be as sensible. Approaches which assume such priors are called Bayesian approaches.

Note that knowing these two things is equivalent to knowing $p(x, \omega)$, but since some approaches only assume knowledge of $p(x|\omega)$ we prefer to separate them.

The goal of decision theory is to decide how to act given a measurement x . We are interested in a **decision function** (or a strategy) $\delta : X \rightarrow A$. There are several criteria for deciding what the best strategy is. Below we focus on the Bayesian approach.

1.1 Bayesian Decision Theory

How do we choose a “good” decision function? Had we known what the state of the world ω is, there would be no problem. Clearly we would just choose the action α that minimized the cost $\lambda(\alpha|\omega)$. But we don’t. The state of the world is **unobserved**.

If we take action α every time we see an observation x we will incur a different cost every time, since ω may have different values every such time (it might rain sometimes when it’s cloudy, and sometimes it might not). Fortunately, we can calculate the **expected cost** of taking action α for a given observation x .

$$R(\alpha|x) = \sum_{\omega} p(\omega|x)\lambda(\alpha|\omega) \quad (1)$$

This is called the **conditional risk**.

Note that in the discrete case this is equivalent to matrix multiplication ($R = \Lambda p(\omega)$)

The probability $p(\omega|x)$ is called the **posterior probability** and is calculated via Bayes rule:

$$p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)} = \frac{p(x|\omega)p(\omega)}{\sum_{\omega} p(x|\omega)p(\omega)} \quad (2)$$

Before seeing x we have the prior distribution over $p(\omega)$ which then becomes the posterior after seeing the observation.

The risk in Equation ?? is defined per value of observation. We would like a decision rule for all values of x so we would like to consider the risk when averaged over all values of x , namely the **overall risk**:

$$R[\delta(x)] = \sum_x R(\delta(x)|x)p(x) = \sum_{x,\omega} p(x, \omega)\lambda(\delta(x)|\omega) \quad (3)$$

The optimal Bayes decision rule is defined as:

$$\delta^*(x) = \arg \min_{\delta(x)} R[\delta(x)] \quad (4)$$

Proposition 1. *The optimal strategy for observation x is to choose*

$$\delta^*(x) = \arg \min_{\alpha} R(\alpha|x) \quad (5)$$

Proof. Consider any other decision rule $\delta(x)$. Then

$$R[\delta(x)] = \sum_x R(\delta(x)|x)p(x) \geq \sum_x R(\delta^*(x)|x)p(x) = R[\delta^*(x)] \quad (6)$$

so that the risk for $\delta^*(x)$ is lower than that of any other decision rule. \square

Definition 1. *The risk $R[\delta^*(x)]$ is called the **Bayes Risk**.*

Definition 2. *Define the **state conditional risk** as the average risk of rule $\delta(x)$ given that the world state is ω :*

$$R(\delta(x)|\omega) = \sum_x p(x|\omega)\lambda(\delta(x)|\omega) \quad (7)$$

Then it's easy to see that

$$R[\delta(x)] = \sum_{\omega} p(\omega)R(\delta(x)|\omega) \quad (8)$$

1.1.1 The two category case

Suppose there are two actions $A = \{\alpha_1, \alpha_2\}$ and two states $\Omega = \{\omega_1, \omega_2\}$. What is the optimal decision rule? Lets write the conditional risk.

$$\begin{aligned} R(\alpha_1|x) &= \lambda(\alpha_1|\omega_1)p(\omega_1|x) + \lambda(\alpha_1|\omega_2)p(\omega_2|x) \\ R(\alpha_2|x) &= \lambda(\alpha_2|\omega_1)p(\omega_1|x) + \lambda(\alpha_2|\omega_2)p(\omega_2|x) \end{aligned}$$

We will take action α_1 if $R(\alpha_1|x) \leq R(\alpha_2|x)$. This is equivalent to:

$$p(\omega_1|x) [\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)] \geq p(\omega_2|x) [\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)] \quad (9)$$

Use the fact that

$$p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)} \quad (10)$$

To get:

$$p(x|\omega_1)p(\omega_1) [\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)] \geq p(x|\omega_2)p(\omega_2) [\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)] \quad (11)$$

Assume wlog that $\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1) \geq 0$ (meaning if the state is ω_1 we should prefer to predict α_1). Then the rule is:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \geq \frac{p(\omega_2)}{p(\omega_1)} \cdot \frac{\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)} \quad (12)$$

The above is called a **likelihood ratio**. Note that the RHS is independent of x . Why does it make sense: suppose ω_2 is a priori much more likely than ω_1 . Then x needs to be such that it's much more likely given ω_1 for us to decide in favor of ω_1 .

Q: What happens at equality?

Show examples with Gaussians.

1.1.2 Error minimization

Suppose we have n states and our action is to predict the correct value of the state. So:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

The conditional risk for observation x is

$$R(\alpha_i|x) = \sum_{\omega} p(\omega|x)\lambda(\alpha|\omega) = \sum_{j \neq i} p(\omega_j|x) = 1 - p(\omega_i|x) \quad (13)$$

Minimizing the risk is equivalent to maximizing $p(\omega_i|x)$ so in this case we should choose the action α_{i^*} where

$$i^* = \arg \max_i p(\omega_i|x) \quad (14)$$

More simply, we will predict the state to be the one that maximizes $p(\omega|x)$

1.1.3 Error minimization: Two states

Assume two states. Then we will predict state ω_1 if

$$p(\omega_0|x) \geq p(\omega_1|x) \quad (15)$$

or

$$\frac{p(x|\omega_0)}{p(x|\omega_1)} \geq \frac{p(\omega_1)}{p(\omega_0)} \quad (16)$$

This partitions the space into x for which we will decide 0 and those that will decide 1. Denote by \mathcal{D}_0 the set of x for which we decide 0.

Example: assume $p(x|\omega_i)$ are two one dimensional Gaussians with same variance σ^2 and mean μ_i . Then:

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu_i)^2} \quad (17)$$

Assume equal priors. Then $x \in \mathcal{D}_1$ if:

$$\begin{aligned} \log p(x|\omega_0) - \log p(x|\omega_1) &\geq 0 \\ -(x - \mu_0)^2 + (x - \mu_1)^2 &\geq 0 \\ 2x(\mu_0 - \mu_1) &\geq \mu_0^2 - \mu_1^2 \\ x &\geq \frac{\mu_0 + \mu_1}{2} \end{aligned}$$

1.1.4 Overall risk. Error minimization in Two states

The expected value of the risk is the probability of error of our procedure. It can be written as:

$$R[\delta(x)] = \sum_{\omega} p(\omega)R(\delta(x)|\omega) = p(\omega_0)R(\delta(x)|\omega_0) + p(\omega_1)R(\delta(x)|\omega_1)$$

$$R(\delta(x)|\omega_0) = \sum_x p(x|\omega_0)\lambda(\delta(x)|\omega_0) = \sum_{x \in \mathcal{D}_1} p(x|\omega_0) = p(x \in \mathcal{D}_1|\omega_0) \quad (18)$$

Similarly $R(\delta(x)|\omega_1) = p(x \in \mathcal{D}_0|\omega_1)$. So:

$$R[\delta(x)] = p(\omega_0)p(x \in \mathcal{D}_1|\omega_0) + p(\omega_1)p(x \in \mathcal{D}_0|\omega_1) \quad (19)$$

2 Hypothesis Testing

In the previous class we learned how to predict the state of the world Ω if we have a prior probability defined on it. What if we want to test a scientific hypothesis (e.g., *is the world round*). It's not clear that a prior is well defined here. The school of thought which avoid assuming priors in such cases is the frequentist approach.

Assume the world states Ω are partitioned into two sets Ω_0 and Ω_1 . Define two hypotheses:

- The **Null Hypothesis** H_0 which corresponds to the hypothesis that $\omega \in \Omega_0$.
- The **Alternative Hypothesis** H_1 which corresponds to the hypothesis that $\omega \in \Omega_1$.

As before we assume we have access to distributions $p(x|\omega)$ for all ω . A test is a procedure that receives x , checks if $x \in \mathcal{D}_1$ (the rejection region or critical region). If it is, we reject H_0 (accept H_1). Otherwise we accept H_0 .

In what sense is such a procedure good or bad? Begin with a simpler case where there is one world state in Ω_0 and one in Ω_1 (denote by ω_1 and ω_2). This is called the two **simple hypotheses** case.

We define two types of errors:

- **Type I error** (e_I or $\alpha(\delta)$ or **false positive**): The probability of rejecting H_0 when in fact it is true (i.e., the world state is ω_0). This is given by $p(x \in \mathcal{D}_1|\omega_0)$. It is also called its size. A test with size at most α is said to have significance level α .
- **Type II error** (e_{II} or $\beta(\delta)$ or **false negative**): The probability of accepting H_0 when in fact it is false (i.e., the world state is ω_1). This is given by $p(x \notin \mathcal{D}_1|\omega_1)$. Note that $1 - \beta$ is sometimes called the power of the test.

Recall from our Bayesian discussions: if we had priors on Ω , we could have defined a procedure that rejects H_0 if

$$\frac{p(x|\omega_1)}{p(x|\omega_0)} \geq \frac{p(\omega_0)}{p(\omega_1)} \quad (20)$$

Since it is the optimal Bayesian strategy we know that it minimizes the risk

$$p(\omega_0)\alpha(\delta) + p(\omega_1)\beta(\delta) \quad (21)$$

Suppose are interested in a test with significance level α . If we never reject then we have it but clearly that's not very good and our test is not sensitive. Also, there are probably tests which *do* reject and still achieve this significance level. Which one is preferable?

Lemma 1. (Neyman and Pearson, 1933) Assume we have a test of the form: Choose ω_1 (i.e. reject the Null) if

$$\frac{p(x|\omega_1)}{p(x|\omega_0)} \geq \theta . \quad (22)$$

Denote this test by $\delta_\theta(x)$. Assume it has type one error $\alpha(\delta_\theta)$ (i.e., $p(x \in \mathcal{D}_1|\omega_0) = \alpha(\delta_\theta)$) and denote its type two error by $\beta(\delta_\theta)$. Then for any other test δ with $\alpha(\delta) \leq \alpha(\delta_\theta)$ it holds that $\beta(\delta) \geq \beta(\delta_\theta)$. Furthermore, if $\alpha(\delta) < \alpha(\delta_\theta)$ then $\beta(\delta) > \beta(\delta_\theta)$.

Proof. Denote the above test by δ_θ . We can define $p(\omega_0), p(\omega_1)$ such that their ratio is exactly θ :

$$\frac{p(\omega_0)}{1 - p(\omega_0)} = \theta$$

$$p(\omega_0) = \frac{\theta}{1 + \theta}, \quad p(\omega_1) = \frac{1}{1 + \theta}$$

Thus the test above is Bayes optimal for this prior. This implies that:

$$p(\omega_0)\alpha(\delta_\theta) + p(\omega_1)\beta(\delta_\theta) \leq p(\omega_0)\alpha(\delta) + p(\omega_1)\beta(\delta) \quad (23)$$

for all other tests δ . Write it as:

$$p(\omega_0)(\alpha(\delta_\theta) - \alpha(\delta)) \leq p(\omega_1)(\beta(\delta) - \beta(\delta_\theta)) \quad (24)$$

And so if $\alpha(\delta_\theta) \geq \alpha(\delta)$ necessarily $\beta(\delta) \geq \beta(\delta_\theta)$ (the second part of the lemma also follows easily). \square

How do we choose θ in practice? Usually we have some α' error that we want to achieve, so we would like to set θ so that the test δ_θ has this error. This means that θ should satisfy:

$$\alpha' = p(x \in D_1(\delta_\theta) | \omega_0) = \int_{x \in D_1(\delta_\theta)} p(x) dx \quad (25)$$

Note that in many cases, there is no need to optimize over θ but rather to consider all possible decision regions D_1 and choose the one that satisfies the above.

As an example, say $p(x|\omega_i)$ is a Gaussian with mean $\mu_0 = 0, \mu_1 = 1$ and $\sigma^2 = 1$. Recall that in this case, the likelihood ratio test is equivalent to decide ω_1 when $x \geq \gamma$ where γ is some constant. So, we want to find a γ such that:

$$\int_{\gamma}^{\infty} \mathcal{N}(x, 0, 1) dx = \alpha' \quad (26)$$

And this can be solved for α using the inverf function. Denote by $\Phi(\gamma)$ the tail integral of a standard Normal distribution between $-\infty$ and γ . Then $\gamma = \Phi^{-1}(1 - \alpha')$.

3 Multi Dimensional Observations

A typical setting is where before we decide how to act we have multiple observations. For example

- say there are two possible coins and we obtain results for n throws before deciding which coin it was.

- Make repeated measurements with a thermometer

Assuming all measurements were made with the same state of the world, and are independent we have:

$$p(x_1, \dots, x_n | \omega) = \prod_i p(x_i | \omega) \quad (27)$$

We call them Independently Identically Distributed (IID).

We will be interested in the behavior of measures based on such samples as n grows.

Suppose we perform a likelihood ratio test. So decide ω_1 if:

$$\frac{p(\omega_1 | x_1, \dots, x_n)}{p(\omega_0 | x_1, \dots, x_n)} \geq t \quad (28)$$

This is equivalent to

$$\begin{aligned} \frac{1}{n} \log \frac{p(\omega_1 | x_1, \dots, x_n)}{p(\omega_0 | x_1, \dots, x_n)} &\geq \frac{1}{n} \log t \\ \frac{1}{n} \log \frac{p(\omega_1)}{p(\omega_0)} + \frac{1}{n} \log \frac{p(x_1, \dots, x_n | \omega_1)}{p(x_1, \dots, x_n | \omega_0)} &\geq \frac{1}{n} \log t \end{aligned}$$

What does this number look like as $n \rightarrow \infty$? That depends on where the x_i are coming from.

$$\frac{1}{n} \log \frac{p(x_1, \dots, x_n | \omega_1)}{p(x_1, \dots, x_n | \omega_0)} = \frac{1}{n} \sum_i \log \frac{p(x_i | \omega_1)}{p(x_i | \omega_0)} \quad (29)$$

This is an empirical average of IID samples from a variable with value $\sum_i \log \frac{p(x_i | \omega_1)}{p(x_i | \omega_0)}$, it thus converges to:

$$\sum_x p(x | \omega_1) \log \frac{p(x | \omega_1)}{p(x | \omega_0)} \equiv D_{KL}[p(x | \omega_1) | p(x | \omega_0)] \quad (30)$$

Or:

$$\sum_x p(x | \omega_0) \log \frac{p(x | \omega_1)}{p(x | \omega_0)} \equiv -D_{KL}[p(x | \omega_0) | p(x | \omega_1)] \quad (31)$$

Similarly the expected value of $\log \frac{p(\omega_1 | x_1, \dots, x_n)}{p(\omega_0 | x_1, \dots, x_n)}$ is

$$E_{p(x_1, \dots, x_n | \omega_1)} \left(\log \frac{p(\omega_1 | x_1, \dots, x_n)}{p(\omega_0 | x_1, \dots, x_n)} \right) = n D_{KL}[p(x | \omega_1) | p(x | \omega_0)] + \log \frac{p(\omega_1)}{p(\omega_0)} \quad (32)$$

$$E_{p(x_1, \dots, x_n | \omega_0)} \left(\log \frac{p(\omega_1 | x_1, \dots, x_n)}{p(\omega_0 | x_1, \dots, x_n)} \right) = -n D_{KL}[p(x | \omega_0) | p(x | \omega_1)] + \log \frac{p(\omega_1)}{p(\omega_0)} \quad (33)$$

As $n \rightarrow \infty$: if the samples are coming from ω_1 we will decide ω_1 if

$$D_{KL}[p(x|\omega_1)|p(x|\omega_0)] \geq 0 \quad (34)$$

meaning always, and therefore we will not make any mistakes in this case. As $n \rightarrow \infty$: if the samples are coming from ω_0 we will decide ω_1 if

$$-D_{KL}[p(x|\omega_0)|p(x|\omega_1)] \geq 0 \quad (35)$$

which is never, so we will never make a mistake in this case as well.

4 Kullback Leibler divergence

The Kullback-Leibler divergence between two distribution is defined as:

$$D_{KL}[p(x)|q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (36)$$

Note that here if $p(x) = 0$ we add 0 regardless of the value of $q(x)$.

It is

- Non negative and is zero iff $p(x) = q(x)$ (in the continuous case it may be zero if p and q are equal almost everywhere).

Proof: Use the fact that $\log x \leq x - 1$:

$$-D_{KL}[p(x)|q(x)] = \sum p(x) \log \frac{q(x)}{p(x)} \leq \sum p(x) \left(1 - \frac{q(x)}{p(x)}\right) = 0 \quad (37)$$

Equality holds iff $\frac{q(x)}{p(x)} = 1$ for all x .

- Non symmetric. For example if there is an x for which $p(x) > 0, q(x) = 0$. Then we will have a divergence of ∞ . This makes sense because it means we are looking at the likelihood ration between $p(x)$ and $q(x)$ observe something that is infinitely more likely under $p(x)$.

5 Sequential Decision Making

Say we first observe x_1 then x_2 etc. Now we wish to decide about ω at every step. At every step we want to decide according to the MAP. Initially we decide based on x_1 by maximizing $p(\omega|x_1)$

$$p(\omega|x_1) = \frac{p(x_1|\omega)p(\omega)}{p(x_1)} \propto p(x_1|\omega)p(\omega) \quad (38)$$

Next:

$$p(\omega|x_1, x_2) = \frac{p(x_1, x_2|\omega)p(\omega)}{p(x_1, x_2)} = \frac{p(x_1|\omega)p(x_2|\omega)p(\omega)}{p(x_1, x_2)} \propto p(x_2|\omega)p(\omega|x_1) \quad (39)$$

$$p(\omega|x_1, x_2, x_3) \propto p(x_1, x_2|\omega)p(x_3|\omega)p(\omega) \propto p(x_3|\omega)p(\omega|x_1, x_2) \quad (40)$$

If we have two states, what happens to the prior? Converges to one for the true state.