DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 2: Bayesian and Frequentist Decision Making

Lecturer: Michael I. Jordan

1 Introduction to Hypothesis Testing and Decision Theory

A **hypothesis test** is a form of statistical inference, typically used to compare two alternatives, or one alternative to a fixed baseline. For example, we may use a hypothesis test to answer questions like: "Is drug A more effective than drug B in reducing blood pressure?" or "Is the recovery time from disease X after taking drug A less than one week?"

We assume that the reality is in one of two states - null, which we will denote 0, or alternative, which we will denote 1. Similarly, the outcome of our test, or our **decision**, will also take on one of those two values.

Example 1.1. *Suppose we know that, with no treatment, the average recovery time from disease X is seven days, and suppose we have a data set of patient records in which patients were treated with drug A after observing that they have disease X. We want to know if taking drug A makes the recovery time longer or shorter, compared to the recovery time when no medication is taken. There is a true reality that says whether the recovery time is really longer or shorter than seven days after taking drug A.*

The following table shows different relations between our decision and the ground-truth.

		decision	
		null (0)	*non-null (1)*
reality	*null* (0)	true negative	false positive
	non-null (1)	false negative	true positive

Table 1.1: Different ground-truth and decision relationships.

Note that here a positive instance corresponds to the *non-null* status, much like in the language of medical diagnoses. When the decision matches reality, we call this a true positive (if the decision was correctly 1), or a true negative (if the decision was correctly 0). Otherwise, if the decision does not match reality, we call this a false negative (if the decision was wrongly 0) or a false positive (if the decision was wrongly 1). For each of the following examples, which errors should we care more about: false positives or false negatives?

- commerce (no fraud (0) vs fraud (1))
- medical (no disease detected (0) vs disease detected (1))
- physics (no Higgs boson found (0) vs Higgs boson found (1))

- search engine (search result not relevant (0) vs search result relevant (1))
- self-driving car (no pedestrian detected (0) vs pedestrian detected (1))

Suppose our binary decisions are classifications, and we want to find a good linear decision boundary for the data below. Everything above the line will be classified as 1, and everything below the line will be classified as 0. In practice, data can be high-dimensional and noisy; as a result, it is likely that any classifier could make mistakes. If we are in a setting where we care a lot about false positives, we could reduce the number of false positives by pushing the decision boundary up (so that we classify more instances as 0). Notice that this immediately implies that we will be making more false negatives. As a conclusion, there is typically a trade-off between false positives and false negatives, and the problem setting specifics should determine which error is more severe.

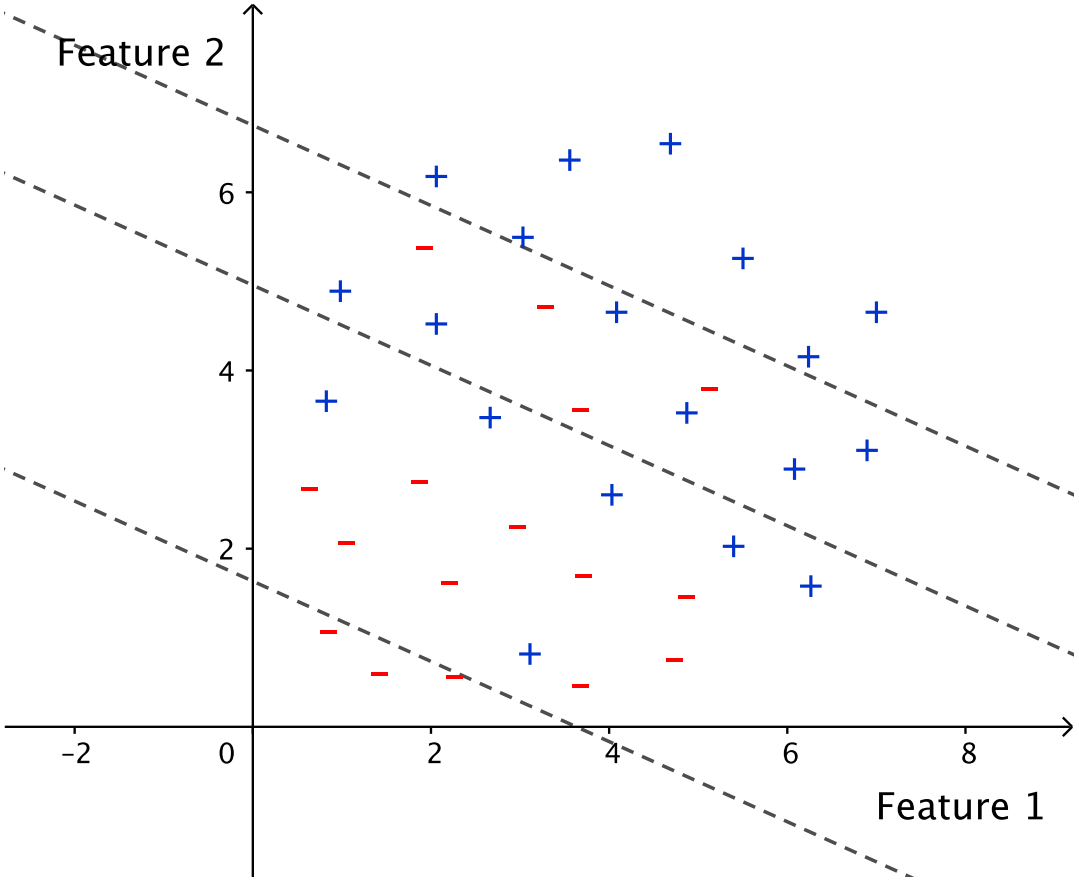

Figure 1.1: Binary classification.

2 Multiple Decisions

Suppose now that we want to make N decisions. For example, we might want to test N different treatments for a given disease, or if we are a company we might test if there is fraud for N different online transactions. We can then talk about the number of true positives, true negatives, false positives, and false negatives as counts, as in Table 1.2.

		decision		
		null (0)	*non-null (1)*	
reality	*null*	n_{00}	n_{01}	$n_{00} + n_{01}$
	non-null	n_{10}	n_{11}	$n_{10} + n_{11}$
		$n_{00} + n_{10}$	$n_{01} + n_{11}$	N

Table 1.2: Different ground truth and decision relationships in multiple testing.

There are different relevant performance criteria we will look at. For example, one is **sensitivity** (aka true positive rate (TPR), power):

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}}.$$

Guaranteeing high sensitivity means that we are discovering almost everything we would ideally like to discover. In other words, this metric quantifies whether our procedure has a good power of detection of interesting signals.

We could look at a similar quantity, but for the upper row instead of the bottom one. This gives us **specificity** (aka true negative rate (TNR)):

$$\text{specificity} = \frac{n_{00}}{n_{00} + n_{10}}.$$

This criterion implies that we don't discover most nulls. Notice that there is a trade-off between sensitivity and specificity. It is easy to have high sensitivity by proclaiming every test a discovery; however, if we also want to ensure high specificity, this is a bad decision rule. Similarly, making no discoveries results in high specificity, but in low sensitivity. Therefore, we ideally want some compromise between the two.

Classical statistical and engineering disciplines typically focus on these "row-wise" quantities. Intuitively, one can think of them as being estimates of relevant conditional probabilities:

$$\text{sensitivity} \approx \mathbb{P}(\text{decision} = 1 \mid \text{reality} = 1), \quad \text{specificity} \approx \mathbb{P}(\text{decision} = 0 \mid \text{reality} = 0).$$

Notice that these quantities do not depend on the prevalence, which is the probability that the reality is 0 (and hence one minus the probability that the reality is 1).

The traditional Neyman-Pearson hypothesis testing framework solves the trade-off between sensitivity and specificity by constructing a constrained optimization problem. The basic idea is that one should set a **significance level** α, and seek to maximize sensitivity, while ensuring that specificity is at least $1 - \alpha$. This is probably the treatment of hypothesis testing you have seen in previous classes.

3 Bayesian and Frequentist Thinking

In statistics, Bayesian and frequentist thinking comprise two major approaches to making decisions. Here we illustrate some of the key differences between the two.

In frequentism, we want to design a procedure that works "on average", or with high probability. We assume that we don't have just one data set, but rather we repeatedly draw data sets independently from a population.

Frequentist hypothesis testing is captured by the Neyman-Pearson framework. There is an unknown ground truth which says whether the hypothesis is null (0) or non-null (1), and for a given significance level α (typically 0.05), the test should guarantee that a discovery is proclaimed with probability at most α, if the ground-truth reality is 0. This setting requires knowing what data we expect to see, if the reality is 0 or 1, respectively.

In Bayesian hypothesis testing, one additionally specifies with what probability the null and non-null occur. This allows us to compute the probability that a hypothesis is 0 (or 1) given the data. We declare a discovery if $\mathbb{P}(H = 0 \mid \text{data})$ is small enough.

Therefore, the frequentist perspective is an unconditional one, requiring inferential procedures to give good answers in repeated use. The Bayesian perspective is a conditional one, making inferences conditional on observed data, as opposed to all possible data one could have observed.

The Bayesian perspective is natural in the setting of a long-term project with a domain expert. For example, if one has a limited data set and works on a multi-year project with a biologist who has good prior knowledge biological phenomena, Bayesian thinking is a sensible approach. The frequentist approach is more "robust", and as such is natural in settings where we develop procedures that will be used repeatedly, in many different, possibly unexpected settings. One example is writing software that will be used by many people for various problems.

Denote by $H \in \{0, 1\}$ the state of reality, and by $D \in \{0, 1\}$ our decision. In frequentist hypothesis testing, we want to guarantee that $\mathbb{P}(D = 1 \mid H = 0)$ is small. In Bayesian hypothesis testing, we apply Bayes' theorem to obtain:

$$\mathbb{P}(H = 0 \mid D = 1) = \frac{\mathbb{P}(D = 1 \mid H = 0)\pi_0}{\mathbb{P}(D = 1)}, \tag{1.1}$$

where we denote by π_0 the prior probability of the hypothesis being null, $\mathbb{P}(H = 0)$. The choice of π_0 is a modeling assumption. Notice that we are implicitly conditioning on the data, because our decisions are direct functions of the data that we input into our inference procedure.

4 Elements of Decision Theory

For the most part, concepts introduced in this section apply to both Bayesians and frequentists.

Denote by X a data set, and let it be generated from some distribution P_θ which is indexed by a ground-truth parameter θ. We define a procedure $\delta(X)$ that operates on the data to make a

decision. Typically, $\delta(X)$ is trying to "guess" θ from X. To measure how good the guess of our procedure is, we take a loss function $\ell(\theta, \delta(X))$, which takes as input the ground-truth parameter, as well as our prediction.

Example 1.2. In the hypothesis testing framework we have discussed so far $\theta \in \{0, 1\}$, $\delta(X) \in \{0, 1\}$. In this case the loss could simply be the 0/1 loss, $\ell(\theta, \delta(X)) = \mathbf{1}\{\theta \neq \delta(X)\}$.

Example 1.3. In many applications, θ can actually take values in a continuous set, for example $\theta \in \mathbb{R}$, in which case our predictions are also continuous $\delta(X) \in \mathbb{R}$. In such a setting it makes sense to pick the loss to be the squared loss $\ell(\theta, \delta(X)) = (\delta(X) - \theta)^2$.

When we design our inference procedure $\delta(\cdot)$, both arguments of the loss function are unknown; θ because we don't know the ground truth, and $\delta(X)$ because we still don't have the data. The treatment of uncertainty regarding θ and X differs between Bayesian and frequentist approaches.

The frequentist approach assumes θ is deterministic, however unknown. In this view, we incur some average error for each of the possible "realities", where the average is taken over the randomness in X, given θ. The frequentist risk is defined as:

$$R(\theta) = \mathbb{E}[\ell(\theta, \delta(X))],$$

where the randomness in the argument of the expectation comes *only* from X.

In the Bayesian view, the parameter θ is also random, and we are interested in the average error given the data X. Here, the average is taken over the randomness in θ. This gives the Bayesian posterior risk:

$$\rho(X) = \mathbb{E}[\ell(\theta, \delta(X))|X].$$

If we accept the Bayesian view of making θ random, then averaging the frequentist risk over θ, and averaging the Bayesian posterior risk over X gives the same number by Fubini's theorem:

$$\mathbb{E}[R(\theta)] = \mathbb{E}[\rho(X)].$$

This number is called the Bayes risk.

5 False Discovery Proportion and False Omission Rate

Let's go back to Table 1.2. As argued before, the row-wise quantities of sensitivity and specificity are predominantly frequentist. A more Bayesian approach would suggest looking at column-wise quantities, because conditioning on a column is the same as conditioning on a decision, which is in turn equivalent to conditioning on having seen certain data.

One such column-wise quantity is the false discovery proportion (FDP):

$$\text{FDP} = \frac{n_{01}}{n_{01} + n_{11}}.$$

This quantifies what fraction of the discoveries declared are actually false. Its column-wise complement is the false omission rate (FOR), defined as:

$$\text{FOR} = \frac{n_{10}}{n_{00} + n_{10}},$$

which quantifies what proportion of all the times we haven't made a discovery there was actually a discovery to be made.

Like the row-oriented quantities from before, these column-oriented fractions can be thought of as estimates of conditional probabilities:

$$\text{FDP} \approx \mathbb{P}(\text{reality} = 0 | \text{decision} = 1), \quad \text{FOR} \approx \mathbb{P}(\text{reality} = 0 | \text{decision} = 1).$$

Unlike sensitivity and specificity, the two quantities above depend on the prevalence (probabilities that the reality is 0 or 1). To see this, rewrite the conditional probability expressions above using equation (1.1).

We care about the FDP when we suspect that for most of our tests, the reality is null, meaning there is no interesting discovery to be made. And indeed, this is the case in most modern testing applications; most tests in large-scale experimentation are a shot in the dark. Moreover, many hypotheses are artificially generated even without any prior supporting evidence, simply because we have great computational power and data resources.