STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 14

A brief recap

- Inference using Bayes' Theorem allows us to summarize our knowledge about unknown parameters after observing the data probabilistically (the ideal inference).
- But it requires treating the parameters as random variables and represent our *a priori* knowledge probabilistically as well.
- Sampling theory doesn't require this, but inference based on it must give up the ideal, and target at more limited inferential goals.
- We have looked at one particularly important example of such goals—the (point and interval) *estimation* problem.
- In the next few weeks we will focus on another important example—the *hypothesis testing* problem.

What is hypothesis testing?

- Estimation is about pinning down the underlying values of unknown parameters from a potentially *large (even infinite) number* of possibilities.
- In the simplest hypothesis testing setting, we ask a *dichotomous* question: If the unknown parameter can take two different values (or two different sets of values), which of the two should be inferred from the data?
- In other words, among two *hypotheses* about the parameter value, which should be accepted and which rejected?
- Let us look at three examples to have a feel of how such problems may arise.

Example I: Zip code recognition

- The parameter θ here is the underlying true number from "0" to "9".
- The data is a 200×300 black/white pixels. How many possible values can the data be?
- $2^{200\times300}$ possible values!
- A probability model is a specification of the probability distribution $p(x|\theta)$ on these $2^{200\times300}$ values.

Testing a hypothesis

- For a single observed pattern X that could have arisen from either "0" or "6", which should be assigned? (Draw a figure.)
- Under a Bayesian perspective, this decision seems straightforward. Let us assign prior probability on θ to characterize our *a priori* knowledge about its value.
- Apply Bayes' theorem to get the posterior probabilities

$$P(\theta = \text{“0”}|x) \quad \text{and} \quad P(\theta = \text{“6”}|x).$$

- Based on this probabilistic summary an assignment can be made.
- More generally, introduce a loss function, $L(\theta, a)$ and Bayes rule is

$$\delta^*(x) = \text{argmin}_a \mathrm{E}\left(L(\theta, a)\,|\,x\right).$$

  Verify: the above rule corresponds to the 0-1 loss $L(\theta, a) = \mathbf{1}(\theta = a)$.
- A Bayesian hypothesis testing problem takes exactly the same form as an estimation problem!

What if we want to adopt the sampling viewpoint?

- We can only use the model: $p(x|\theta)$, not the prior.
- One possibility: Assign $\theta =$"0" or "6" based on whether $p(x|\theta) > 0.5$.
- This is problematic as $p(x|\theta)$ may be >0.5 (or <0.5) for both "0" and "6"!
- How about assign "0" or "6" based on whether

  $$p(x|\theta = \text{“0”}) > p(x|\theta = \text{“6”})?$$

  In other words, let us *compare the likelihood* at the two θ values and pick the one with larger likelihood.
- If we repeat the experiment many times, what would happen if $\theta = 0$ versus if $\theta = 6$.
- The idea of comparing the likelihoods under the two hypotheses seems to make sense intuitively.

Example II: Testing the quality of a lot

A computer company purchases many different electronic components from suppliers for making a laptop.

- Want to make sure that the components are of good quality in terms of life time.
- The quality of the components in each lot is usually similar, and one can model the life-time X of a component from a given lot to be an Exponential(λ) random variable. So the expected life-time for the components in this lot is $E(X) = 1/\lambda$.
- Suppose a lot is considered good if $\lambda \leq 1.0$ and bad if $\lambda > 1.0$.
- How do we judge whether a lot is good or bad?
- Note that in this problem we are deciding between two *sets* of values for λ, instead of two specific values as in the previous example.

- Obviously we can't test every single component in a lot.
- A common strategy is to randomly *sample* a few, say n, components from the lot, and measure their life-time

$$X_1, X_2, \ldots, X_n.$$

- How do we judge whether this lot is good or bad based on these life-times?
- An intuitive idea: How about we calculate the sample average life-time $\bar{X}$ and see whether it is large or small?
- Is it enough to just compare $\bar{X}$ with 1.0?
- *If we repeat the experiment many times*, what would happen if H_0 is true versus if H_1 is true?

Example III: Contingency tables

- In the 1880's, Francis Galton carried out a study on whether men and women choose their spouse on the basis of height.
- Do tall men tend to marry tall women, do short women tend to marry short men, or do people choose spouses regardless of each other's height?
- He was studying the heritability of height, and would like to know how he should take the correlation between parental heights into the analysis.
- He collected the following data set.

Galton's data

		Wife:		
		Tall	Medium	Short
	Tall	18	28	14
Husband:	Medium	20	51	28
	Short	12	25	9

- Question: Do these data support or refute the *hypothesis of independence* of spouses' heights?
- What is the sampling model here? What are the unknown parameters? How is the hypothesis formulated in terms of the unknown parameters?

Testing simple hypotheses

- Suppose we observe data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from some distribution $f(\mathbf{x}|\theta)$ or $p(\mathbf{x}|\theta)$.
- What is a *simple hypothesis*?
- A *simple hypothesis* is one that completely specifies the sampling distribution of the data $f(\mathbf{x}|\theta)$.
- For example, in the zipcode recognition example, if we know what

$$f(\mathbf{x}|\theta = \text{“0”}) \quad \text{and} \quad f(\mathbf{x}|\theta = \text{“6”})$$

are exactly then $\theta =$“0” and $\theta =$“6” are both simple hypotheses.

- In contrast, in the lot testing example, $\lambda \leq 1.0$ and $\lambda > 1.0$ do not completely specifies the distribution of $f(\mathbf{x}|\lambda)$.
- This kind of hypotheses, which specify the distribtion $f(\mathbf{x}|\lambda)$ to be one of a collection of probability distributions, are called *composite* hypotheses.
- We will get to testing composite hypotheses later.
- Is $\lambda = 2.5$ is a simple or composite hypothesis?

Another example of simple versus composite hypotheses.

- Suppose our data are i.i.d. observations from a $N(\mu, \sigma^2)$ distribution.
- First consider the case when σ^2 is known.
- Is $\mu = 5$ a simple or composite hypothesis?
- Now what if σ^2 is unknown?

How about the the example of testing independence on a contingency table?

How do we compare two simple hypotheses

- This is the simplest situation for hypothesis testing.
- The two simple hypotheses can be formally treated as

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

  where each of θ_0 and θ_1 completely specifies $f(\mathbf{x}|\theta)$.
- We make a choice between these two possibilities based on the data $\mathbf{X}$ we observe.
- Let us consider this as a binary decision problem—reject or accept H_0.
- When $\mathbf{X}$ take values in a set of possible values $\mathscr{R}$, we will *reject* θ_0 and choose θ_1.
- Otherwise, we accept, or do not reject, θ_0.

The rejection region

- Therefore a test, i.e. a *decision rule* for choosing one of the two hypotheses based on the data, can be specified by the rejection region $\mathscr{R}$.

$$\begin{aligned} &\text{If } \mathbf{X} \text{ is not in the rejection region } \mathscr{R} \text{ decide } \theta = \theta_0. \\ &\text{If } \mathbf{X} \text{ is in the rejection region } \mathscr{R} \text{ decide } \theta = \theta_1. \end{aligned}$$

- So constructing a test boils down to choosing the corresponding rejection region $\mathscr{R}$. That is, choosing the data values corresponding to each decision.

Examples of rejection regions

- Zipcode example. Consider a 200×300 black/white grid. Total of 2^{60000} possible $\mathbf{X}$ values.
- There are numerous—$2^{2^{60000}}$—different ways to define a rejection region.
- (Draw a figure.)

Examples of rejection regions

- Consider the lot testing example, where the data are continuous.
- The sample space (i.e. the collection of possible data values) is infinite.
- Infinite number of possible rejection regions.
- In particular, one can define rejection regions based on common statistics—such as sample mean, sample median and sample variance.

- This same formulation applies to testing composite hypotheses too.
- For the lot testing example, one possible rejection region may be

$$\mathscr{R} = \{(X_1, X_2, \ldots, X_n) : \;\; \bar{X} < 1.0.\}$$

Another may be that

$$\mathscr{R} = \{(X_1, X_2, \ldots, X_n) : \;\; \text{sample median}(X) < 1.0.\}$$

Yet another may be that

$$\mathscr{R} = \{(X_1, X_2, \ldots, X_n) : \;\; \text{sample variance}(X) < 1.0.\}$$

The question is

- How do we choose the rejection region?
- It turns out that there is a general procedure with which one can construct good tests.
- The idea is related to the maximum likelihood principle we used for the estimation problem.
- Let's start with the testing of simple vs simple hypothesis.

Constructing good tests: comparing likelihoods

- How about we compare the likelihoods under the two hypotheses.
- Choose the hypothesis with the higher likelihood. That is

$$\mathscr{R}: \quad \text{Those } \mathbf{x} \text{ values such that } \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > 1.$$

- The "best" test is very much along this line!

The likelihood ratio test

- The *likelihood ratio statistic*

$$\frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)}$$

  measures the relative evidence for the data under the two hypotheses.
- A rejection region based on this statistic is

$$\frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)} > K$$

  for some constant K.
- Tests with rejection regions of this form are called *likelihood ratio tests*.
- It turns out that these tests are the "best" test to use (under certain criteria).
- Next time we will introduce a few notions to make precise what "good", and "best" mean.