

---

## Reasoning Versus Learning, Point Estimation

Lecture #1: Tuesday, 4 January 2005  
Lecturer: Prof. Charles Elkan  
Scribe: Daniel Hsu  
Reviewer: Samory Kpotufe

### 1 Reasoning versus learning

This course is about statistical learning, or machine learning using statistics. Statistics is based on probability theory, which in turn is a branch of mathematics based on measure theory. Though statistics and probability theory are related, they differ from one another in very important ways.

Probability theory is deductive; it is used for reasoning. Deductive reasoning is powerful in that the conclusions drawn using deduction are guaranteed to be true when the premises are true. A typical problem from probability theory is: “Given a certain probability distribution (typically a pdf, i.e. a probability density function) (e.g. the Gaussian pdf  $D = N(0, 4)$ ), what is the probability of a certain event (e.g.  $X > 10$  where the random variable  $X \sim D$ )?”

Deductive reasoning cannot go from the particular to the general. This limitation is inherited by probability theory. For example, even if we observe that all students in many classes were born in the 20th century, deductive reasoning cannot conclude that all students in all classes were born in the 20th century.

Statistical inference is inductive as opposed to deductive; it is used for learning. The power of inductive inference is the ability to draw general conclusions from observations, just as a child or an agent learns from experience. A typical problem of statistical inference is: “Suppose we assume a family of pdfs (e.g. the Gaussian family of distributions  $N(\mu, \sigma^2)$ ), and we observe certain outcomes (e.g.  $x_1 = 1.3, x_2 = 0.7, x_3 = 0.8, x_4 = 0.2$ ), then what is the best pdf in the family to use for future reasoning?” In general, we make certain background assumptions, we see a set of training data, and we want to infer something general. The learned information is often something that can be used for reasoning with probability theory.

The conclusions made using inductive learning are not guaranteed to be true. In the example problem above, we cannot determine with certainty the probability distribution that governs the experiment. Also, we almost always have to make assumptions about the training data (e.g. assume the data is normal) that cannot be checked for correctness. A lot of research explores which assumptions are made in natural (human) learning.

**Example 1 (Neuron population).** Consider a large population of neurons. At age  $t$ , a proportion  $\pi(t)$  are still alive. For  $i \in \{1, 2, \dots, k\}$ , we take a random sample of  $n$  neurons at age  $t_i$  (assume the samples are independent) and observe that  $r_i$  are still alive. Let  $x = \langle r_1, r_2, \dots, r_k \rangle$ . We will consider both a reasoning problem and a learning problem.

A reasoning problem assumes  $\pi$  is a known function of  $t$  and asks for  $P(x)$ . Since the samples are independent,

$$P(x) = \prod_{i=1}^k P(r_i)$$

where

$$P(r_i) = \binom{n}{r_i} \pi(t_i)^{r_i} (1 - \pi(t_i))^{n-r_i}.$$

Note that we make the assumption that the survival of each neuron is independent of the survival of each other neuron. We observe that  $r_i$  follows the binomial distribution with parameter  $\pi(t_i)$ , so the task is the same as computing the probability of  $r_i$  heads on  $n$  biased coin tosses.

But  $\pi$  is usually not known! A learning problem is to guess the function  $\pi$ . A reasonable monotonicity assumption is that  $\pi$  is a non-increasing function on  $[0, 1]$ . We make a guess, or estimate, for  $\pi$ . An obvious estimate is  $\hat{\pi}(t_i) = \frac{r_i}{n}$  for each  $i$ . Note that even if the monotonicity assumption is true, it might be violated by the estimate (e.g. for  $n = 10, r_1 = 5, r_2 = 6$ , we have  $\hat{\pi}(t_1) = \frac{1}{2} < \frac{3}{5} = \hat{\pi}(t_2)$ ). However, we don't conclude that the assumption is false! This is a result of random variation in the data: the result of observations will be different for different training sets.

Note the difference between an *estimator* and an *estimate*. An estimator is a learning algorithm (e.g. formula for  $\hat{\pi}(t_i)$ ), whereas an estimate is the result of applying an estimator to a particular training set. For example, we could propose another estimator (this one will ensure monotonicity in the estimate), given by the following algorithm:

start with  $a_i = \frac{r_i}{n}$

repeat

while there exists  $i$  such that  $a_i < a_{i+1}$

$$\text{padding-left: 80px; } b := \frac{a_i + a_{i+1}}{2}$$

$$\text{padding-left: 80px; } a_i := b$$

$$\text{padding-left: 80px; } a_{i+1} := b$$

until for all  $i$ ,  $a_i \geq a_{i+1}$

## 2 Point estimation

The neuron example concerns estimating a whole probability distribution (in fact, a different whole distribution for each age  $t$ ). But some applications just require a *point estimate*, the estimate of a single real number (e.g. the mean of a distribution). Generally speaking, it should be easier to estimate the value of just one number rather than a whole distribution.

The general estimation problem assumes a family  $\{P_\theta \mid \theta \in \Theta\}$  of pdfs on a sample space  $\mathcal{X} = \{x\}$ , where  $x$  is one training set. Note, each  $\theta$  (e.g.  $\theta = \langle \mu, \sigma^2 \rangle$ ) gives a different  $P(x|\theta)$ , and for each  $\theta$ ,  $\int_{\mathcal{X}} P(x|\theta)dx = 1$ . The task is to find a good estimator, i.e. a function  $f : \mathcal{X} \rightarrow \Theta$ .

One possible meaning of “good” is an estimator that yields  $\hat{\theta}$  so that  $P(x|\hat{\theta})$  is high. We will return to this idea, called *maximum likelihood*, later in the course.

The point estimation problem has the same assumption of  $P_\theta$ , but we are only interested in some property of the distribution, not necessarily the distribution itself. For example, let  $g : \Theta \rightarrow \mathbb{R}$  (e.g.  $g(\langle \mu, \sigma^2 \rangle) = \mu$ ,  $g(\theta) = \mathbb{E}[x|x \sim P_\theta]$ ). The task is to find a good estimator  $h : \mathcal{X} \rightarrow \mathbb{R}$  so that  $h(x)$  is close to  $g(\theta)$  for all  $x$  and  $\theta$ .

**Example 2.** In Example 1, the general estimation problem is to estimate  $\pi(t)$  for all  $t$ . A point problem is to estimate  $t^*$  such that  $\pi(t^*) = 0.5$ , i.e. the half-life of the population.

**Example 3.** Suppose  $x = \langle x_1, x_2, \dots, x_n \rangle$  is an independent and identically distributed (i.i.d.) sample with each  $x_i$  drawn from  $N(\mu, \sigma^2)$ . Let  $g(\langle \mu, \sigma^2 \rangle) = \mu$  (we only care about the mean). An obvious estimator for  $g$  is given by

$$h(x) = \frac{1}{n} \sum_{i=1}^n x_i,$$

which is the empirical average of the sample.