Statistical Approaches to
Learning and Discovery

Week 1: Some Basic Concepts from Statistics and Information Theory

January 15, 2003

Today's Agenda

- High-level view
- Sufficient statistics
- Data processing inequality (no free statistical lunch)
- Estimators: Bias, variance, Cramér-Rao
- Exponential families

Supervised vs. Unsupervised Learning

Have a sequence (or set) of inputs $x_1, x_2, x_3, \ldots,$, "naturally" occuring, collected by hand, or generated by machine

Supervised Learning: Machine is given desired outputs $y_1, y_2, y_3, \ldots,$ and goal is to learn to produce the correct output given a new input. This doesn't specify how "correct" should be assessed... Distinction between classification (discrete y_i) and regression (continuous y_i).

Unsupervised Learning: Goal is to build representations of x that can be used for reasoning, decision making, predicting, communicating, etc. Task is often not well specified.

Supervised vs. Unsupervised Learning (cont.).

Semi-Supervised: Same as supervised, but some of the values y_i are missing in the training set, and the unlabeled $x_i's$ are incorporated.

Inference vs. Learning

Estimation/Learning: Selecting parameters, a distribution over parameters, or a set of cdf's for a statistical problem based on data.

Inference: Making predictions, computing statistics, expectations, or marginal probabilities for a statistical model that has already been estimated/learned.

Parameters

A statistical family with a finite collection of adjustable parameters is the starting point for a *parametric* estimation problem.

If there are an infinite number of adjustable parameters–typically entire functions or cdf's, then the problem (or approach) is said to be *non-parametric.*

Parametric vs. Non-Parametric

This can be confusing, since often "non-parametric" problems seem to have many more "parameters" than a typical parametric problem.

Non-parametric approaches make fewer assumptions about the form or "shape" of the distribution being estimated.

However, the distinction is sometimes subtle (e.g., neural nets)

A Simple Estimator

Suppose that $X_1, X_2, \dots, X_n \sim \mathcal{N}(\theta, 1)$ (iid).

We want to determine θ from the sample. Two options:

1. X_1, since clearly $E(X_1) = \theta$
2. $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Also mean θ

Which is better? Well, depends what "good" means. In fact, $\overline{X}_n$ is the minimum mean squared error unbiased estimator.

Role of computation is not emphasized in classical statistics...

Sufficiency

Suppose $X_i \sim f(\cdot \,|\, \theta)$, for $\theta \in \Theta \subset \mathbb{R}^m$.

A *statistic* is just a function of the sample: $T(X_1, \ldots, X_n)$. *It's a random variable.*

Supose there is a statistician and a computer scientist. The statistician has all of the data $X_1, \ldots, X_n$. The computer scientist only keeps a "hash" of the data $T(X_1, \ldots, X_n)$.

Who can make better estimates of θ, or in general make better inferences?

Sufficiency (cont.)

In general, the statistician can do better, but if T is a *sufficient statistic* then the computer scientist will be able to do just as well.

In this case, intuitively, $T(X_1, \ldots, X_n)$ contains all of the "information" in the sample about θ, and the individual values are irrelevant.

(We'll give a precise meaning to this later...)

Example 1: Bernoulli

$X_1, X_2, \ldots, X_n$ are n coin tosses. $X_i \sim$ Bernoulli(θ).

Given n, the number of "heads" is a sufficient statistic for θ.

$$Pr(X_i = x_i \,|\, n, T(X) = k) \;=\; \begin{cases} \frac{1}{\binom{n}{k}} & \text{if } \sum_i x_i = k \\ 0 & \text{otherwise} \end{cases}$$

More generally, for a multinomial $\theta = (p_1, p_2, \ldots, p_t)$, the vector of counts $(n_1, \ldots, n_t)$ is sufficient, where $n_j = \sum_{i=1}^{n} \delta(x_j = i)$.

Example 2: Gaussian

Take

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} = \mathcal{N}(\theta, 1)$$

A sufficient statistic is $\overline{X}_n = \frac{1}{n}\sum_i X_i$.

$\overline{X}_n$ and $\frac{1}{n}\sum_i (X_i - \overline{X}_n)^2$ are sufficient for μ and σ^2 if $\theta = (\mu, \sigma^2)$.

Example 2: Uniform

Take

$$X_i \sim \text{Uniform}(0, \theta)$$

A sufficient statistic is $T(X_1, \ldots, X_n) = \max_i X_i$.

Neyman Factorization Criterion

A statistic $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint pdf can be factored as

$$f_n(\boldsymbol{x} \,|\, \theta) = u(\boldsymbol{x})\, v(T((x), \theta))$$

Information

Now let's go back and give a precise meaning to "all of the relevant information about θ is in the sufficient statistic"

So far, we've only been thinking of $\mathcal{X}_i$ as random, not θ. We'll now need to treat θ as a random variable.

Data Processing Inequality

"No clever manipulation of the data can improve the inferences that can be made from the data."

Note: this is a statement about statistics, not computation

Information Theory Concepts

For a discrete distribution $p_1, p_2, \ldots, p_n$, or random variable X with $p(X = x_i) = p_i$, *entropy*

$$H(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

in *bits* of information.

Conditional entropy $H(X \mid Y)$ is

$$\begin{aligned} H(X \mid Y) &= \sum_y p(Y = y)\, H(X \mid Y = y) \\ &= -\sum_y p(y) \sum_x p(x \mid y) \log_2 p(x \mid y) \end{aligned}$$

Information Theory Concepts (cont.)

Mutual information $I(X; Y)$

$$\begin{aligned} I(X; Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \end{aligned}$$

Informally, "the average value of a hint." Amount by which knowing X reduces the average code length needed to compress Y.

Markov Chains

$X \longrightarrow Y \longrightarrow Z$ forms a *Markov chain* in case the conditional distribution of Z is independent of X.

Equivalently, in case X and Z are conditionally independent given Y. Note: *"time" symmetric*

(Concept extends to spatial processes, or "random fields")

Data Processing Inequality

If $X \longrightarrow Y \longrightarrow Z$ is a Markov chain, then

$$I(X; Y) \geq I(X; Z)$$

In particular, since $X \longrightarrow Y \longrightarrow g(Y)$,

$$I(X; Y) \geq I(X; g(Y))$$

Sufficiency Revisited

Since $\Theta \longrightarrow X \longrightarrow T(X)$ is a Markov chain, we have that $I(\Theta; X) \geq I(\Theta; T(X))$.

However, if $\Theta \longrightarrow T(X) \longrightarrow X$ is a Markov chain also, i.e., *$T(X)$ is sufficient*, then we have equality:

$$I(\Theta; T(X)) = I(\Theta; X)$$

(Historical note: Notion of sufficiency due to Fisher; Formulation in terms of mutual information due to Kullback.)

Estimation: Basic Concepts

Point estimation: choose a *single* parameter $\hat{\theta}$ or cdf, or other prediction.

Note: $\hat{\theta}$ is a random variable, since it is a function of the data (which is random):

$$\hat{\theta}_n = g(X_1, X_2, \ldots, X_n)$$

where g *represents an algorithm* for computing the point estimate.

Bias

The *bias* of a point estimator is

$$\text{bias}(\hat{\theta}_n) = E_F[\hat{\theta}_n] - \theta$$

An estimator is *unbiased* if

$$E_F[\hat{\theta}_n] = \theta$$

where $X_1, X_2, \ldots, X_n$ are iid $\sim F$.

Consistency

A point estimate of a parameter θ is *consistent* if

$$\hat{\theta}_n \longrightarrow \theta \quad \text{(in probability)}$$

The *standard error* is the standard deviation of $\hat{\theta}_n$:

$$\text{se}(\hat{\theta}_n) = \sqrt{E_F(\hat{\theta}_n - E_F(\hat{\theta}_n))^2}$$

For an unbiased estimator this is

$$\text{se}(\hat{\theta}_n) = \sqrt{E_F(\hat{\theta}_n - \theta)^2}$$

Note that since the expectation is the "true" expectation over the data, this is in general impossible to compute.

Example

Let $X_i \sim$ Bernoulli(θ).

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

satisfies

$$E[\hat{\theta}_n] = \frac{1}{n} \cdot n\theta = \theta$$

so this is an unbiased estimate of θ.

Example (cont.)

The standard error is

$$\begin{aligned} \text{se}(\hat{\theta}_n) &= \sqrt{E\left(\left(\frac{1}{n}\sum X_i\right)^2 - \theta^2\right)} \\ &= \sqrt{\frac{\theta(1-\theta)}{n}} \end{aligned}$$

and so can't be computed. The estimated standard error is

$$\hat{\text{se}} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

Mean Squared Error

The *mean squared error* (MSE) of an estimator is

$$E[(\hat{\theta}_n - \theta)^2]$$

Another way of looking at this is

$$\begin{aligned} MSE &= E[(\hat{\theta}_n - \theta)^2] \\ &= E[((\hat{\theta}_n - E[\hat{\theta}_n])^2 + (E[\hat{\theta}_n] - \theta))^2] \\ &= \text{Var}(\hat{\theta}_n) + \text{bias}^2(\hat{\theta}_n) \end{aligned}$$

Fundamental tradeoff.

Asymptotically Normal

An estimator is *asymptotically normal* in case

$$\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \rightsquigarrow \mathcal{N}(0,1)$$

Point Estimation for Parametric Families

We have a family $\mathcal{F} = \{f_\theta(x), \theta \in \Theta\}$ and want to estimate certain parameters of interest.

Maximum Likelihood

The most commonly used method for point estimation. Given a family $\mathcal{F} = \{f(x \,|\, \theta)\}$ and data $X_1, X_2, \ldots, X_n$, the *likelihood function* is defined as

$$\mathcal{L}_n(\theta) = \prod_i f(X_i \,|\, \theta)$$

and the *log-likelihood function* is given by

$$\begin{aligned} \ell_n(\theta) &= \log \mathcal{L}_n(\theta) \\ &= \sum_i \log f(X_i \,|\, \theta) \end{aligned}$$

Maximum Likelihood

The *maximum likelihood estimator* is

$$\hat{\theta} = \text{argmax}_\Theta \ell_n(\theta)$$

(whenever this exists)

What is the Best Possible Estimator?

What is the minimum variance of an (unbiased) estimator of θ?

Take $f(x \,|\, \theta)$ where $\theta \in \mathbb{R}$ (1-dimensional for simplicity).

Let's look at the change in log-likelihood as a function of θ. The *score* $s(X, \theta)$ is defined as

$$s(X, \theta) = \frac{\partial}{\partial \theta} \log f(X \,|\, \theta)$$

This has mean zero (with respect to $f(\cdot \,|\, \theta)$)

Fisher Information and Cramér-Rao

Fisher information is the variance of the score:

$$\begin{aligned} J(\theta) &= E_\theta(s^2) \\ &= E_\theta\left(\frac{\partial}{\partial\theta}\log f(X\,|\,\theta)\right)^2 \end{aligned}$$

Basic additivity property: The Fisher information of n iid samples is $nJ(\theta)$.

Cramér-Rao Inequality: The mean-squared error of any unbiased estimator $T(X)$ for θ satisfies

$$E_\theta(T-\theta)^2 = \text{Var}(T) \geq \frac{1}{J(\theta)}$$

Example: Gaussian

Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ where σ is known.

It's easy to compute that $J(\theta) = \frac{1}{\sigma^2}$.

The sample mean meets the Cramér-Rao lower bound:

$$E_\theta(\overline{X}_n - \theta)^2 = \frac{\sigma^2}{n} = \frac{1}{J_n(\theta)}$$

It is an *efficient estimator*

Asymptotic Normality of the MLE

Under some regularity conditions, the MLE is asymptotically normal, with standard error given by the inverse Fisher information:

$$\frac{(\widehat{\theta}-\theta)}{\sqrt{1/nJ(\theta)}} \rightsquigarrow \mathcal{N}(0,1)$$

This enables us to compute asymptotic confidence intervals

Different Emphasis for Estimation/Learning

Traditional Statistics	*Machine Learning*
consistency	computational efficiency
bias	statistical efficiency
statistical efficiency	bias
computational efficiency	consistency