Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

22. November 2022

Universität Potsdam

Data reduction

Task: given a set of samples $\{x_1, \ldots, x_n\}$ estimate a parameter θ of an associate density $f(x; \theta)$

Example: given n iid samples $x_i \sim \mathcal{N}(\theta, 1)$, estimate θ

Idea: extract key features that summarizes the information contained in a sample, e.g., in the form of a statistic T (function defined on set of samples)

Example: $T(x_1, \ldots, x_n) = x_1 + \cdots + x_n$
Note: T should capture all relevant information to infer unknown parameter, i.e, sample sets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ with $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$ should lead to the same estimated parameter $\hat{\theta}$

Sufficiency

Def: A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample $x_1, \dots, x_n$ given the value $T(x_1, \dots, x_n)$ does not depend on θ.

Example: Let $X_1, \dots, X_n$ be iid Bernoulli random variables with parameter θ, $0 < \theta < 1$. Then $T(X_1, \dots, X_n) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. As $T(X_1, \dots, X_n) \sim \text{binomial}(n, \theta)$ the ratio of the pmfs is

$$\frac{p(x_1, \dots, x_n|\theta)}{q(T(x_1, \dots, x_n)|\theta)} = \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \quad (\text{define } t = \sum x_i) \tag{1}$$

$$= \frac{\theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \tag{2}$$

$$= \frac{\theta^t(1-\theta)^{(n-t)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\sum x_i}} \tag{3}$$

Sufficient statistic

Theorem: If $p(x_1, \ldots, x_n|\theta)$ is the joint pdf or pmf of $X_1, \ldots, X_n$, and q$(t\theta)$ is the pdf or pmf of $T(X_1, \ldots, X_n)$, then $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if, and only if, for every $x_1, \ldots, x_n$ in the sample space the ratio $p(x_1, \ldots, x_n|\theta)/q(T(x_1, \ldots, x_n)|\theta)$ is constant as a function of θ.

Proof: Must verify that for any fixed values of $x_1, \ldots, x_n$ and t, the conditional probability $\mathbb{P}(X_1, \ldots, X_n = x_1, \ldots, x_n|T(X_1, \ldots, X_n) = t)$ is the same for all values of θ. This probability is zero for all values of θ if $T(x_1, \ldots, x_n) \neq t$. So the case that remains to be checked is

$$
\begin{aligned}
&\mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n|T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)) \\
&\quad= \frac{\mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n \text{ and } T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n))}{P_\theta(T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n))} \\
&\quad= \frac{\mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n)}{P_\theta(T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n))} = \frac{p(x_1, \ldots, x_n|\theta)}{q(T(x_1, \ldots, x_n)|\theta)}
\end{aligned}
$$

$\square$

Factorization Theorem

Theorem: Let $f(x_1, \ldots, x_n)$ denote the joint pdf of the associated random variables $X_1, \ldots, X_n$. A statistic $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x_1, \ldots, x_n)$ such that, for all sample points $x_1, \ldots, x_n$ and all parameter points θ

$$f(x_1, \ldots, x_n)|\theta) = g(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n) \tag{4}$$

Proof (only for discrete distibutions): Suppose $T(X_1, \ldots, X_n)$ is a sufficient statistic. Choose $g(T(x_1, \ldots, x_n)) = \mathbb{P}_\theta(T(X_1, \ldots, X_n) = t)$ and $h(x_1, \ldots, x_n) = \mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n | T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n))$. Then

$$\begin{aligned}
f(x_1, \ldots, x_n)|\theta) &= \mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n) \\
&= \mathbb{P}_\theta(X_1, \ldots, X_n = x_1, \ldots, x_n \text{ and } T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)) \\
&= \mathbb{P}_\theta(T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)) \\
&\times \mathbb{P}(X_1, \ldots, X_n = x_1, \ldots, x_n | T(X_1, \ldots, X_n) = T(x_1, \ldots, x_n)) \text{ (sufficiency)} \\
&= g(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n)
\end{aligned}$$

$\square$

Example

For the normal model where θ is the mean and $T(x_1, \ldots, x_n) = \bar{x}$

$$f(x_1, \ldots, x_n) = (2\pi)^{-n/2} \exp(-\sum_{i=1}^{n}(x_i - \theta)^2/(2\sigma^2)) \tag{5}$$

and

$$h(x_1, \ldots, x_n) = (2\pi)^{-n/2} \exp(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)) \tag{6}$$

So we have

$$g(t|\theta) = exp(-n(t - \theta)^2/(2\sigma^2)) \tag{7}$$

Hypothesis Testing

Def: A hypothesis is a statement about a population parameter θ.

Example: A statement about the range of the unknown mean θ of a considered distribution (for instance a Gaussian distribution $\mathcal{N}(\theta, 1)$)

Def: The complementary hypotheses in a hypothesis testing problem is called the *null hypothesis* and the *alternative hypothesis*. They are denotes by H_0 and H_1.

Hypothesis Testing

Def: A hypothesis testing procedure or hypothesis test is a rule that specifies:

1. For which sample value the decision is made to accept H_0 as true.
2. For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region.

Various testing ideas

Likelihood Ratio Tests

Reminder: For a given set of iid samples $x_1, \ldots, x_n$ from random variables X_i distributed according to density $f(x|\theta)$ the likelihood with respect to parameter θ is

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) \tag{8}$$

Def: The likelihood ratio test (LRT) statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\Theta_0} L(\theta|x_1, \ldots, x_n)}{\sup_{\Theta} L(\theta|x_1, \ldots, x_n)}. \tag{9}$$

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{x : \lambda(x) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

Example: Normal LRT

Example: For a given set of iid samples $x_1, \ldots, x_n \sim \mathcal{N}(\theta, 1)$

$$\lambda(x_1, \ldots, x_n) = \frac{(2\pi)^{-n/2} \exp(-\sum_{i=1}^n (x_i - \theta_0)^2/2)}{(2\pi)^{-n/2} \exp(-\sum_{i=1}^n (x_i - \bar{x})^2/2)} \tag{10}$$

$$= \exp\left[\Big(-\sum_{i=1}^n (x_i - \theta_0)^2 + \sum_{i=1}^n (x_i - \bar{x})^2\Big)/2\right] \tag{11}$$

$$\begin{aligned}
\sum_{i=1}^n (x_i - \theta_0)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \theta_0)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \theta_0) \sum_{i=1}^n (x_i - \bar{x}) + \sum_{i=1}^n (\bar{x} - \theta_0)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \theta_0) \underbrace{\Big(\sum_{i=1}^n (x_i) - n\bar{x}\Big)}_{=0} + \sum_{i=1}^n (\bar{x} - \theta_0)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2
\end{aligned}$$

Example: Normal LRT

Inserting

$$\sum_{i=1}^{n}(x_i - \theta_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2$$

in

$$\begin{aligned} \lambda(x_1, \dots, x_n) &= \exp\left[\Big(- \sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\Big)/2\right] \\ &= \exp\left[\Big(- n(\bar{x} - \theta_0)^2\Big)/2\right] \end{aligned}$$

Ansatz: An LRT test rejects H_0 for small values of $\lambda(x_1, \dots, x_n)$. Using the rejection region

$$\{x_1, \dots, x_n : \lambda(x) \geq c\} = \{x_1, \dots, x_n : |\bar{x} - \theta_0| \geq \sqrt{-2(\log c)/n}\} \qquad (12)$$

Sufficient statistic

Theorem: If $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(x_1, \ldots, x_n)$ are the LRT statistics based on T and $X_1, \ldots, X_n$, respectively, then $\lambda^*(T(x_1, \ldots, x_n)) = \lambda(x_1, \ldots, x_n)$ for every $x_1, \ldots, x_n$ in the sample space.

Proof: From the Factorization theorem, we know we can write the pdf or pmf of $X_1, \ldots, X_n$ as $f(x_1, \ldots, x_n|\theta) = g(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n)$ where $g(T(x_1, \ldots, x_n))$ is the pdf or pmf of T and $h(x_1, \ldots, x_n)$ does not depend on θ. Thus

$$\begin{aligned}
\lambda(x_1, \ldots, x_n) &= \frac{\sup_{\Theta_0} L(\theta|x_1, \ldots, x_n)}{\sup_{\Theta} L(\theta|x_1, \ldots, x_n)} = \frac{\sup_{\Theta_0} f(x_1, \ldots, x_n|\theta)}{\sup_{\Theta} f(x_1, \ldots, x_n|\theta)} \\
&= \frac{\sup_{\Theta_0} g(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n)}{\sup_{\Theta} g(T(x_1, \ldots, x_n))h(x_1, \ldots, x_n)} = \frac{\sup_{\Theta_0} g(T(x_1, \ldots, x_n))}{\sup_{\Theta} g(T(x_1, \ldots, x_n))} \\
&= \frac{\sup_{\Theta_0} L(\theta|T(x_1, \ldots, x_n))}{\sup_{\Theta} L(\theta|T(x_1, \ldots, x_n))} = \lambda^*(T(x_1, \ldots, x_n))
\end{aligned}$$

$\square$

Test function

Def: A test function $\phi(x_1, \ldots, x_n)$ for a hypothesis testing procedure is a function on the sample space whose value is one if a sample set $x_1, \ldots, x_n$ is in the rejection area and zero if $x_1, \ldots, x_n$ is in the acceptance region. In other words $\phi(x_1, \ldots, x_n)$ is an indicator function of the rejection region.

Def: A invariant test with respect to a function $g(x_1, \ldots, x_n)$ is any test whose test function satisfies $\phi(x_1, \ldots, x_n) = \phi(g(x_1, \ldots, x_n))$ for any $x_1, \ldots, x_n$ in the sample space.

Invariant hypothesis testing

Assuming: the following models

- the joint family of densities of $X_1, \ldots, X_n$ is $\{f(x_1, \ldots, x_n|\theta) : \theta \in \Theta\}$
- the joint family of densities for $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$ are $\{h(y_1, \ldots, y_n|\theta) : \theta \in \Theta\}$

Def: A hypothesis testing problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is invariant under the transformation $y_1, \ldots, y_n = g(x_1, \ldots, x_n)$ if

- $\{f(x_1, \ldots, x_n|\theta) : \theta \in \Theta_0\} = \{h(y_1, \ldots, y_n|\theta) : \theta \in \Theta_0\}$
- $\{f(x_1, \ldots, x_n|\theta) : \theta \in \Theta_0^c\} = \{h(y_1, \ldots, y_n|\theta) : \theta \in \Theta_0^c\}$

Bayesian tests

Idea: Use Bayesian model, i.e, the posterior $\pi(\theta|x_1,\dots,x_n)$ to define test:

- $P(\theta\in\Theta_0|x_1,\dots,x_n) = P(H_0 \text{ is true}|x_1,\dots,x_n)$
- $P(\theta\in\Theta_0^c|x_1,\dots,x_n) = P(H_1 \text{ is true}|x_1,\dots,x_n)$
- **need:** hypothesis test that is binary (i.e., either true or false)

Def: One hypothesis tests inspired by a Bayesian model is given by setting H_0 as true if $P(\theta\in\Theta_0|X_1,\dots,X_n) \geq \mathrm{P}(\theta\in\Theta_0^c|X_1,\dots,X_n)$.