Stats 210A Lecture Notes

Kurtland Chua

Contents

1 Measure Theory

Definition 1.1. A *σ-field* on X is a collection of subsets of X which must include $\varnothing$ and X that is closed under complements and countable union. ◇

We refer to a space and a σ-field defined on that space as a *measurable space*.

Definition 1.2. Let X be a set. A *measure* on X is a function μ mapping sets within a σ-field on X to $[0, \infty]$ such that the following properties hold:

1. $\mu(\varnothing) = 0$.
2. If $A_1, A_2, \ldots$ are mutually disjoint sets in the σ-field, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$ ◇

Examples of measures include the counting measure on countable sets and the Lebesgue measure on $\mathbb{R}^n$ (n-volume).

Definition 1.3. A *probability measure* is a measure on X such that $\mu(X) = 1$. ◇

Given a measure, we can define integrals that weigh subsets by their measure. For example, for the counting measure μ on X, we have that for any real-valued function f on X:

$$\int_X f \, \mathrm{d}\mu = \sum_{x \in X} f(x)$$

As another example, if μ is the Lebesgue measure on $\mathbb{R}$, then for any real-valued function f:

$$\int_{\mathbb{R}} f \, \mathrm{d}\mu = \int_{-\infty}^{\infty} f(x) \, \mathrm{d}x$$

Definition 1.4. Let $(X, \mathcal{F})$ be a measurable space, and let μ, ν be measures on the space. Then, we say that μ is *absolutely continuous* with respect to ν if $\mu(A) = 0$ whenever $\nu(A) = 0$ for any $A \in \mathcal{F}$. Absolute continuity of measures is often denoted as $\mu \ll \nu$. ◇

If P is a probability measure on X that is absolutely continuous with respect to another measure μ, then we can define a density function f such that

$$P(A) = \int_A f \, \mathrm{d}\mu.$$

The function f, also known as the *Radon-Nikodym derivative*, is often denoted as $\mathrm{d}P/\mathrm{d}\mu$. The existence of such density functions is guaranteed by the Radon-Nikodym theorem. We then have that in general

$$\int_A g \, \mathrm{d}P = \int_A g \left(\frac{\mathrm{d}P}{\mathrm{d}\mu} \right) \, \mathrm{d}\mu.$$

The density of a probability measure with respect to a counting measure is referred to as a *probability mass function*. Similarly, the density of a probability measure with respect to the Lebesgue measure is referred to a *probability density function*. Observe that the Radon-Nikodym derivative in general is not unique, since they are defined up to null sets (with respect to μ).

Definition 1.5. Let $(\Omega, \mathcal{F})$ be a measurable space with a probability measure P. We refer to the space together with the probability measure as a *probability space*. Elements of Ω are then referred to as *outcomes*, and sets in the σ-field are *events*. ◇

Definition 1.6. A *random variable* is a measurable function $X : \Omega \to \mathcal{X}$. We say that X has *distribution* Q (or $X \sim Q$) if

$$P(X \in B) = P(\{w \in \Omega \mid X(w) \in B\}) = Q(B).$$ ◇

Definition 1.7. The *expectation* of $X : \Omega \to \mathcal{X}$ is defined as

$$\mathbb{E}[X] = \int_\Omega X \, \mathrm{d}P = \int_{\mathcal{X}} \mathrm{Id} \, \mathrm{d}Q$$ ◇

2 Estimation

Definition 2.1. A *statistical model* is family of candidate probability distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$. ◇

The goal of estimation is that given some observation $X \sim P_\theta$, where θ is an unknown parameter, we want to estimate $g(\theta)$ for some general function g, called the *estimand*.

Definition 2.2. A *statistic* is any function $T(X)$ of the data. ◇

The estimator $\delta(X)$ of $g(\theta)$ is a statistic.

Definition 2.3. A *loss function* is a function $L(\theta, d)$ measuring how bad an estimate is. $\diamond$

For example, we could use the squared-error loss $L(\theta, d) = \|g(\theta) - d\|^2$. Typically, loss functions are nonnegative, and $L(\theta, g(\theta)) = 0$. However, we typically do not use loss functions on their own.

Definition 2.4. The *risk function* is the expected loss

$$R(\theta, \delta(\cdot)) = \mathbb{E}_\theta \left[L(\theta, \delta(X)) \right] \quad \diamond.$$

If we set L to the squared-error loss, then the corresponding risk function is known as the mean squared error (MSE).

Example 2.1. One might want to estimate the probability p that a biased coin lands heads from the result of flipping the coin n times. Let X be the number of heads after fliiping the coin n times. Assuming that the trials are independent, then $X \sim \text{Bin}(n, \theta)$. Then, the density is given by

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^x$$

A reasonable estimator is $\delta(X) = X/n$. Note that $\mathbb{E}\left[X/n\right] = \theta$, so δ is an unbiased estimator. If we use the MSE as the risk function, then

$$MSE(\theta, \delta(\cdot)) = \text{Var}\left[\frac{X}{n}\right] = \frac{1}{n}\theta(1-\theta).$$

Another estimator that is reasonable is $\delta_2(X) = (X+3)/(X+6)$, which effectively biases the estimate towards $1/2$. $\triangle$

Definition 2.5. An estimator δ is *inadmissible* is there exists another estimator δ^* such that

- $R(\theta, \delta(\cdot)) \geq R(\theta, \delta^*(\cdot))$ for any θ in the parameter space.
- For some θ in the space, then $R(\theta, \delta(\cdot)) > R(\theta, \delta^*(\cdot))$.

Then, we say that δ^* *dominates* δ. $\diamond$

It is not in general possible to find the best estimator that dominates all other estimators. Therefore, there are several strategies we can take.

We can first summarize the risk as a scalar in several ways. For one, we can consider the average-case risk and find

$$\underset{\delta}{\text{argmin}} \int_\Theta R(\theta, \delta) \, \mathrm{d}\Lambda(\theta),$$

where Λ is chosen carefully to represent the average case. The above estimator is also known as the *Bayes estimator*, where Λ represents a prior over the parameter. Another way to summarize the risk is the worst-case risk; we would then find

$$\underset{\delta}{\text{argmin}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

This is also known as a *minimax estimator*.

Besides summarizing the risk as a scalar, we can also constrain the choice of estimator. For one, we can consider only the set of unbiased estimators (i.e. $\mathbb{E}\left[\delta(X)\right] = \theta$ for all θ).

3 Exponential Families

Definition 3.1. An *s-parameter exponential family* is a family of probability densities $\{p_\eta \mid \eta \in \Xi\}$ with respect to a measure μ on $\mathcal{X}$ of the form

$$p_\eta(x) = \exp\left[\eta^T T(x) - A(\eta)\right] h(x),$$

where $T : \mathcal{X} \to \mathbb{R}^s$ is a sufficient statistic, $h : \mathcal{X} \to \mathbb{R}$ is the *carrier density* or *Bayes density*, $\eta \in \Xi \subseteq \mathbb{R}^s$ is the *natural parameter*, and $A : \Xi \to \mathbb{R}$ is the *cumulant-generating function*, or *normalizing constant*. $\diamond$

Observe that the CGF $A(\eta)$ is completely determined by T and h as

$$A(\eta) = \log \int_{\mathcal{X}} \exp\left[\eta^T T(x)\right] h(x) \mathrm{d}\mu(x).$$

We say that p_η is *normalizable* if and only if $A(\eta) < \infty$. The *natural parameter space* Ξ is then defined to be the set $\{\eta \in \mathbb{R}^s \mid A(\eta) < \infty\}$. We then say that the family is in the *canonical form* if we allow η to take any values in the natural parameter space.

Observe that without loss of generality, we can

- Change h to $\tilde{h}(x) = p_0(x)$, assuming that 0 is in the natural parameter space. Then, changing $A(\eta)$ to $\tilde{A}(\eta) = A(\eta) - A(0)$ results in an equivalent family.
- Assume $M \in R^{s \times s}$ is invertible, and define $\tilde{T}(x) = MT(x)$. Then, defining $\tilde{\eta} = (M^{-1})^T \eta$ results in the same family.
- We can also set $h(x) = 1$ for all x if we change μ to $\tilde{\mu}$, chosen so that its Radon-Nikodym derivative of with respect to μ is h.

We can interpret an exponential family as starting with a baseline measure p_0, and then applying an *exponential tilt*, which is multiplication by $\exp\left[\eta^T T(x)\right]$, followed by renormalization.

Note that an exponential tilt cannot change the support of the distribution.

Oftentimes, it is more natural to represent a family using η parametrized by a vector θ so that

$$p_\theta(x) = \exp\left[\eta(\theta)^T T(x) - B(\theta)\right] h(x), \; B(\theta) = A(\eta(\theta)).$$

Example 3.1. Let $X \sim \mathcal{N}[\mu, \sigma^2]$ be a random variable, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Let $\theta = [\mu, \sigma^2]$. Then, we can write

$$p_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]\right\}$$

Therefore, we can write

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \quad B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2).$$

That is, we can view a Gaussian random variable as the normal Lebesgue measure on the real line, tilted using the parameters above. $\triangle$

Most distributions that are used in practice are exponential families.

Another nice property of exponential families is that if we have n i.i.d. samples $x_1, \ldots, x_N$ from a density that is part of an exponential family, the joint distribution is also an exponential family. Suppose the density is given by

$$p_X(x) = \exp\left[-\eta^T T(x) - A(\eta)\right] h(x).$$

Then, the joint density is given by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_N) = \exp\left[-\eta^T \sum_{i=1}^{N} T(x_i) - nA(\eta)\right] \prod_{i=1}^{N} h(x_i)$$

The sufficient statistic $T(x)$ follows a closely related exponential family. Assume that X is a random variable, where $X \sim \exp\left[\eta^T T(x) - A(\eta)\right] h(x)$, where h is discrete. Then,

$$P_\eta(T(x) \in B) = \sum_{x \in T^{-1}(B)} \exp\left[\eta^T T(x) - A(\eta)\right] h(x) = \sum_{t \in B} \exp\left[\eta^T t - A(\eta)\right] \sum_{x \in T^{-1}(t)} h(x)$$

That is, the distribution of $T(X)$ is another exponential family where $T(X)$ is the sufficient statistic.

3.1 Differential Identities

Write

$$e^{A(\eta)} = \int e^{\eta^T T(x)} h(x) \, \mathrm{d}\mu(x).$$

Within the interior of the natural parameter space, we can differentiate under the integral sign to obtain interesting identities. Then,

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \int \frac{\partial}{\partial \eta_j} e^{\eta^T T(x)} h(x) \, \mathrm{d}\mu(x) \implies \left.\frac{\partial A}{\partial \eta_j}\right|_\eta = \int T_j e^{\eta^T T - A(\eta)} h \, \mathrm{d}\mu = \mathbb{E}_\eta\left[T_j(x)\right]$$

Differentiating twice,

$$\frac{\partial^2}{\partial \eta_i \partial \eta_j} e^{A(\eta)} = \int \frac{\partial^2}{\partial \eta_i \partial \eta_j} e^{\eta^T T(x)} h(x) \, \mathrm{d}\mu(x) \implies \frac{\partial^2 A}{\partial \eta_i \partial \eta_j} + \frac{\partial A}{\partial \eta_i}\frac{\partial A}{\partial \eta_j} = \int T_i T_j e^{\eta^T T - A(\eta)} h \, \mathrm{d}\mu.$$

Therefore, using the identity for the first moment, we find that

$$\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = \mathbb{E}_\eta\left[T_i T_j\right] - \mathbb{E}_\eta\left[T_i\right]\mathbb{E}_\eta\left[T_j\right] = \mathrm{Cov}_\eta\left[T_i, T_j\right]$$

In general, the moment generating function of an exponential family is given by

$$M_\eta^T(u) = \exp\left[A(\eta + u) - A(\eta)\right].$$

Therefore, we can obtain n^{th} order moments by differentiating $M_\eta^T(u)$ n times with respect to u at 0.

The cumulant generating function is given by

$$K_\eta^T(u) = \log M_\eta^T(u) = A(\eta + u) - A(\eta).$$

4 Sufficiency

Definition 4.1. Let $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ be a statistical model for data X. Then, $T(X)$ is sufficient for the family $\mathcal{P}$ if $P_\theta(X|T)$ does not depend on θ. $\diamond$

Definition 4.2. Suppose $X_1, \ldots, X_n$ are i.i.d. Bern(θ) random variables. Then,

$$X \sim \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}.$$

Then, we know that $T(X) = \sum_{x_i} \sim \text{Binom}(n, \theta)$, and

$$T(X) \sim \theta^t (1-\theta)^{n-t} \binom{n}{t}$$

Then,

$$P_\theta(X|T) = \frac{\theta^{\sum X_i}(1-\theta)^{\sum X_i} \mathbb{1}\left[\sum X_i = t\right]}{\theta^t (1-\theta)^{n-t}\binom{n}{t}} = \frac{\mathbb{1}\left[\sum X_i = 1\right]}{\binom{n}{t}}$$ $\diamond$

We can interpret sufficiency as first generating T based on the parameter θ, and then generating X that is consistent with T. At the point when X is being generated, θ is ignored since T contains all required information to generate X.

The *sufficiency principle* states that if $T(X)$ is sufficient, then any statistical procedure should only depend on $T(X)$. In particular, if $\tilde{X}$ is data generated using $T(X)$, then X and $\tilde{X}$ are identical in distribution so $\delta(X)$ and $\delta(\tilde{X})$ are also identical in distribution. Therefore, $T(X)$ tells us everything we need to know about X.

Note that sufficiency is always with respect to a model. That is, if we later on decide to expand our model, then the $T(X)$ may no longer be sufficient for the expanded model.

Theorem 4.1 (Factorization)**.** *Let $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ be a family of densities with respect to a common measure μ. Then, $T(X)$ is sufficient if and only if there exists functions $g_\theta, h_\theta \geq 0$ such that p_θ factorizes as $g_\theta(T(X))h(X)$ almost everywhere.*

Example 4.1. By the factorization theorem, it is easy to see that for any s-parameter exponential family, $T(X)$ is sufficient. $\triangle$

Example 4.2. If we have n i.i.d. samples from any statistical model, then the order statistics (i.e. the samples in sorted order) is sufficient for the model. Intuitively, given the order statistics, then we can regenerate the data by uniformly sampling from all possible orderings of the data (because the samples are i.i.d.). $\triangle$

Example 4.3. Assume $X_1, X_2, \ldots, X_n$ are i.i.d. samples from Unif$(\theta, \theta+1)$. Then, the maximum and the minimum of the samples together form the sufficient statistic for the model. $\triangle$

4.1 Minimal Sufficiency

Observe that for n i.i.d. samples from a Gaussian distribution with unknown mean and unit variance, there are multiple possible choices of the sufficient statistics. In particular, we have the sum of the samples, the order statistics, or even just the data itself, and so on. However, note that some statistics can be generated using other statistics. For example, given the order statistics or the data itself, we can exactly recover the sum of the samples, but not the other way around.

Proposition 4.1. *If $T(X)$ is sufficient and $T(X) = f(S(X))$, then $S(X)$ is sufficient.*

The proposition above follows easily from the factorization theorem.

Definition 4.3. A sufficient statistic $T(X)$ is *minimal sufficient* if for every other sufficient statistic $S(X)$, there exists some f such that $T(X) = f(S(X))$. $\diamond$

Intuitively, if $T(X)$ is sufficient, then it is minimally sufficient if it can be recovered from any other sufficient statistic.

The following theorem provides a way to test minimal sufficiency.

Theorem 4.2. *Assume that $\mathcal{P} = \{p_\theta \mid \theta \in \Theta\}$ is a family of densities with respect to a common measure μ. Assume that $T(X)$ is sufficient. If for any x, y, $p_\theta(x) \propto_\theta p_\theta(y)$ implies that $T(x) = T(y)$, then T is minimal sufficient.*

Example 4.4. Consider an s-parameter exponential family parametrized by θ with sufficient statistic $T(X)$. Then, observe that there exists some constant C such that for any θ,

$$p_\theta(x) \propto_\theta p_\theta(y) \iff e^{\eta(\theta)^T T(x)} \propto_\theta e^{\eta(\theta)^T T(y)} \iff \eta(\theta)^T T(x) = \eta(\theta)^T T(y) + C$$

Consequently, for any $\theta_1, \theta_2 \in \Theta$,

$$[\eta(\theta_1) - \eta(\theta_2)]^T [T(x) - T(y)] = 0$$

Therefore, $T(x) - T(y)$ is orthogonal to the span of $\{\eta(\theta_1) - \eta(\theta_2) \mid \theta_1, \theta_2 \in \Theta\}$. It follows from the previous theorem that T is minimal if the aforementioned span is all of $\mathbb{R}^s$. $\triangle$