Advanced Machine Learning **Fall 2020**

Lecture 3: Density Estimation

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes are adapted from ETH's Advanced Machine Learning Course and the book All Of Statistics, Larry Wasserman, Springer.*

3.1 Parametric Inference

We now turn our attention to parametric models, that is, models of the form:

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$$

where the $\theta \subset \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, ..., \theta_k)$ is the parameter. The problem of inference then reduces to the problem of estimating the parameter θ.
Often, we are only interested in some function $T(\theta)$. For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ then $\mu = T(\theta)$ is called the parameter of interest and σ is called a nuisance parameter.

3.1.1 Maximum Likelihood

The most common method for estimating parameters in a parametric model is the maximum likelihood method. Let $X_1, ..., X_n$ be IID with pdf $f(x;\theta)$.

Definition 3.1 *The* ***likelihood function*** *is defined by:*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The ***log-likelihood function*** *is defined by* $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we treat it is a function of the parameter θ. Thus, $\mathcal{L}_n(\theta) : \Theta \to [0, \infty)$. The likelihood function is not a density function: in general, it is not true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

Definition 3.2 *The* ***maximum likelihood estimator*** *MLE, denoted by* $\hat{\theta}_n$*, is the value of* θ *that maximizes* $\mathcal{L}_n(\theta)$.

The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same result as maximizing the likelihood. Often, it is easier to work with the log-likelihood.

Claim 3.3 *If we multiply* $\mathcal{L}_n(\theta)$ *by any positive constant* c *(not depending on* θ*) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.*

Example 3.1 *Let $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:*

$$\begin{aligned}\mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\} \\ &= \sigma^{-n} \exp\{-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2\} \\ &= \sigma^{-n} \exp\{-\frac{nS^2}{2\sigma^2}\} \exp\{-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\}\end{aligned}$$

where $\bar{X} = n^{-1} \sum_i X_i$ is the sample mean and $S^2 = n^{-1} \sum_i (X_i - \bar{X})^2$ .The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$ and then expanding the square. The log-likelihood is:

$$l(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}$$

Solving the equations:

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood.

3.1.2 Properties of Maximum Likelihood Estimators

Under certain conditions on the model, the maximum likelihood estimator $\hat{\theta}_n$ possesses many properties that make it an appealing choice of estimator. The main properties of the MLE are:

1. The MLE is **consistent**: $\hat{\theta}_n \xrightarrow{P} \theta^*$ where θ^* denotes the true value of the parameter θ;
2. The MLE is **equivariant**: if $\hat{\theta}_n$ is the MLE of θ then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$;
3. The MLE is **asymptotically Normal**: $(\hat{\theta} - \theta^*/\hat{\text{se}}) \sim \mathcal{N}(0, 1)$; also, the estimated standard error $\hat{\text{se}}$ can often be computed analytically;
4. The MLE is **asymptotically optimal** or **efficient**: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples. That is, $\hat{\theta}_n$ minimizes $\mathbb{E}\big[(\hat{\theta}_n - \theta^*)^2\big]$ as $n \to \infty$;
5. The MLE is approximately the Bayes estimator.

The properties we discuss only hold if the model satisfies certain regularity conditions. These are essentially smoothness conditions on $f(x; \theta)$, unless otherwise stated we shall tacitly assume that these conditions hold.

3.1.3 Understanding Asymptotic efficiency

The expected square error is a measure for quantifying how good an estimator $\hat{\theta}$ is:

$$\mathbb{E}\big[(\hat{\theta} - \theta_0)^2\big]$$

The Rao-Cramer bound shows that there does not exists an estimator that reaches $\mathbb{E}\big[(\hat{\theta} - \theta_0)^2\big] = 0$

Theorem 3.4 *For any estimator $\hat{\theta}$ of θ it holds that:*

$$\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2\big] \geq \frac{\big(\frac{\partial}{\partial\theta}b_{\hat{\theta}}+1\big)^2}{\mathbb{E}_{x|\theta}\big[\Lambda^2\big]} + b_{\hat{\theta}}^2$$

Where:

$$\Lambda = \frac{\partial}{\partial\theta}\log p(x|\theta) = \frac{1}{p(x|\theta)}\frac{\partial}{\partial\theta}p(x|\theta) \quad \textit{and} \quad b_{\hat{\theta}} = \mathbb{E}_{x|\theta}[\hat{\theta}]-\theta$$

Proof:

$$\mathbb{E}_{x|\theta}[\Lambda] = \int_x p(x|\theta)\Lambda\; dx = \int_x \frac{\partial}{\partial\theta}p(x|\theta)dx = \frac{\partial}{\partial\theta}\overbrace{\int_x p(x|\theta)dx}^{=1} = 0$$

$$\mathbb{E}_{x|\theta}[\Lambda\hat{\theta}] = \int_x p(x|\theta)\Lambda\hat{\theta}\; dx = \int_x \frac{\partial}{\partial\theta}p(x|\theta)\hat{\theta}\; dx = \frac{\partial}{\partial\theta}\int_x p(x|\theta)\hat{\theta}\; dx = \frac{\partial}{\partial\theta}\mathbb{E}_{x|\theta}[\hat{\theta}] = \frac{\partial}{\partial\theta}(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)+1 = \frac{\partial}{\partial\theta}b_{\hat{\theta}}+1$$

Consider the cross-correlation between Λ and $\hat{\theta}$:

$$\Big(\mathbb{E}_{x|\theta}\big[(\Lambda-\overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0})(\hat{\theta}-\mathbb{E}_{x|\theta}[\hat{\theta}])\big]\Big)^2 = \Big(\mathbb{E}_{x|\theta}[\Lambda\hat{\theta}]-\mathbb{E}_{x|\theta}\big[\Lambda\mathbb{E}_{x|\theta}[\hat{\theta}]\big]\Big)^2 = \Big(\mathbb{E}_{x|\theta}[\Lambda\hat{\theta}]-\overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0}\,\mathbb{E}_{x|\theta}[\hat{\theta}]\Big)^2 = \Big(\mathbb{E}_{x|\theta}[\Lambda\hat{\theta}]\Big)^2$$

Now, let's consider Cauchy-Schwarz inequality i.e. $\big(\mathbb{E}[xy]\big)^2 \leq \mathbb{E}[x^2]\mathbb{E}[y^2]$ applied to the cross-correlation:

$$\Big(\mathbb{E}_{x|\theta}\big[(\Lambda-\overbrace{\mathbb{E}_{x|\theta}[\Lambda]}^{=0})(\hat{\theta}-\mathbb{E}_{x|\theta}[\hat{\theta}]\Big)^2 \leq \mathbb{E}_{x|\theta}[\Lambda^2]\;\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\mathbb{E}_{x|\theta}[\hat{\theta}])^2\big] = \mathbb{E}_{x|\theta}[\Lambda^2]\;\mathbb{E}_{x|\theta}\Big[\big((\hat{\theta}-\theta)-(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)\big)^2\Big]$$

$$= \mathbb{E}_{x|\theta}[\Lambda^2]\;\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2+(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)^2-2(\hat{\theta}-\theta)(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)\big]$$

$$= \mathbb{E}_{x|\theta}[\Lambda^2]\;\Big\{\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2\big]+\overbrace{\mathbb{E}_{x|\theta}\big[(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)^2-2(\hat{\theta}-\theta)(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)\big]}^{-b_{\hat{\theta}}^2}\Big\} = \mathbb{E}_{x|\theta}[\Lambda^2]\;\Big\{\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2\big]-b_{\hat{\theta}}^2\Big\}$$

It's easy to verify that $\mathbb{E}_{x|\theta}\big[(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)^2-2(\hat{\theta}-\theta)(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta)\big] = -b_{\hat{\theta}}^2$:

$$\mathbb{E}_{x|\theta}\Big[\mathbb{E}_{x|\theta}^2[\hat{\theta}]+\theta^2\cancel{-2\theta\mathbb{E}_{x|\theta}[\hat{\theta}]}-2\hat{\theta}\mathbb{E}_{x|\theta}[\hat{\theta}]+2\hat{\theta}\theta+\cancel{2\theta\mathbb{E}_{x|\theta}[\hat{\theta}]}-2\theta^2\Big]$$

$$= \mathbb{E}_{x|\theta}^2[\hat{\theta}]+\mathbb{E}_{x|\theta}[\theta^2]-2\mathbb{E}_{x|\theta}^2[\hat{\theta}]+2\theta\mathbb{E}_{x|\theta}[\hat{\theta}]-2\mathbb{E}_{x|\theta}[\theta^2]$$

$$= -\mathbb{E}_{x|\theta}^2[\hat{\theta}]-\mathbb{E}_{x|\theta}[\theta^2]+2\theta\mathbb{E}_{x|\theta}[\hat{\theta}] = -\mathbb{E}_{x|\theta}^2[\hat{\theta}]-\theta^2+2\theta\mathbb{E}_{x|\theta}[\hat{\theta}] = -\big(\mathbb{E}_{x|\theta}[\hat{\theta}]-\theta\big)^2 = -b_{\hat{\theta}}^2$$

Finally, from the inequality proved earlier we know that:

$$\big(\mathbb{E}_{x|\theta}[\Lambda\hat{\theta}]\big)^2 = \big(\frac{\partial}{\partial\theta}b_{\hat{\theta}}+1\big)^2 \leq \mathbb{E}_{x|\theta}[\Lambda^2]\;\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2-b_{\hat{\theta}}^2\big]$$

It follows that:

$$\mathbb{E}_{x|\theta}\big[(\hat{\theta}-\theta)^2\big] \geq \frac{\big(\frac{\partial}{\partial\theta}b_{\hat{\theta}}+1\big)^2}{\mathbb{E}_{x|\theta}\big[\Lambda^2\big]} + b_{\hat{\theta}}^2$$

■

3.1.4 Stein Estimator

For finite samples, the maximum-likelihood estimator is not necessarily efficient.
Consider a multivariate random variable with distribution $\mathcal{N}(\theta_0, \sigma^2 I)$ with range $\mathbb{R}^d$ and $d \geq 3$. If we sample a single point y from this distribution then the Stein Estimator is:

$$\hat{\theta}_{JS} := \Big(1 - \frac{(d-2)\sigma^2}{||y||^2}\Big)y$$

It is possible to prove that the Stein Estimator is better than the maximum-likelihood estimator for any θ_0.
That is:

$$\mathbb{E}\Big[(\hat{\theta}_{JS} - \theta_0)^2\Big] \leq \mathbb{E}\Big[(\hat{\theta}_{ML} - \theta_0)^2\Big] \text{ for any } \theta_0$$

Moreover, the inequality is strict for some values of θ_0.

3.2 Bayesian Learning

Bayesian inference is usually carried out in the following way:

- θ is considered to be a **random variable** with distribution $p(\theta|\mathcal{X})$.
- $X \sim p(x)$ and $p(x)$ is unknown.
- $p(x|\theta)$ is a statistical model that reflects our beliefs about x given θ.

We are looking for $p(X = x|\mathcal{X})$, i.e., the probability of x given the sample set $\mathcal{X}$ (class conditional density):

$$p(X = x|\mathcal{X}) = \int \underbrace{p(x, \theta|\mathcal{X})}_{p(x|\theta,\mathcal{X})p(\theta|\mathcal{X})} d\theta = \int p(x|\theta, \mathcal{X})p(\theta|\mathcal{X})d\theta = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

Where $p(x|\theta, \mathcal{X}) = p(x|\theta)$ since $x_i \in \mathcal{X}$ and x are i.i.d.

Moreover, asymptotically it holds that $p(\theta|\mathcal{X}) \sim \delta(\theta - \hat{\theta})$; intuitively, this follows from the fact that $\hat{\theta} \xrightarrow{p} \theta_{true}$.
Thus, in the asymptotic case, we can approximate the integral with:

$$p(X = x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta \approx \int p(x|\theta)\delta(\theta - \hat{\theta})d\theta = p(x|\hat{\theta})$$

This approximation was used in the early days of Bayesian inference when it was not possible to evaluate the integral.

3.2.1 Bayesian Learning of a Normal Distribution

Let us begin with a simple example in which we consider a single Gaussian random variable x. We shall suppose that the variance σ^2 is known, and we consider the task of inferring the mean μ given a set of N observations:

- The likelihood is $p(x|\mu) = \mathcal{N}(\mu, \sigma^2)$
- The prior is $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$
- The data is $\mathcal{X} = \{x_1, \ldots, x_n\}$

We want to compute the posterior distribution $p(\mu|\mathcal{X})$:

$$p(\mu|\mathcal{X}) \propto p(\mathcal{X}|\mu)p(\mu) \implies p(\mu|\mathcal{X}) = \alpha\, p(\mathcal{X}|\mu)p(\mu) = \alpha \cdot \prod_{i \le n} \Big\{ \frac{1}{\sqrt{2\pi\sigma}} \exp\Big(-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2 \Big) \Big\} \cdot \frac{1}{\sqrt{2\pi\sigma_0}} \exp\Big(-\frac{1}{2}(\frac{\mu - \mu_0}{\sigma_0})^2 \Big)$$

$$= \alpha' \cdot \prod_{i \le n} \Big\{ \exp\Big(-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2 \Big) \Big\} \cdot \exp\Big(-\frac{1}{2}(\frac{\mu - \mu_0}{\sigma_0})^2 \Big) = \alpha' \cdot \exp\Big\{ -\frac{1}{2}\Big(\sum_{i \le n} \Big(\frac{x_i - \mu}{\sigma}\Big)^2 + \Big(\frac{\mu - \mu_0}{\sigma_0}\Big)^2 \Big) \Big\}$$

Expanding the squares we get:

$$p(\mu|\mathcal{X}) = \alpha' \cdot \exp\Big(\mu^2 \overbrace{(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})}^{a} - 2\mu \overbrace{(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{i \le n} x_i^2)}^{b} + c \Big)$$

Which we know is a Gaussian Distribution, i.e. $p(\mu|\mathcal{X}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, because the exponent is a quadratic form. Furthermore, by completing the square we know that:

$$\mu_n = \frac{b}{a} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

$$\sigma_n^2 = \frac{1}{a} = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

It is worth spending a moment studying the form of the posterior mean and variance. First of all, note that the mean of the posterior is a compromise between μ_0 and the maximum likelihood solution $\hat{\mu}$. If the number of observed data points $n = 0$ then μ_n reduces to the prior mean as expected. For $n \to \infty$, μ_n is given by the maximum likelihood solution.
Similarly, consider the result for the variance of the posterior distribution σ_n^2. With no observed data points, we have the prior variance, whereas if the number of data points $n \to \infty$, the variance goes to zero and the posterior distribution becomes infinitely peaked around the maximum likelihood solution.
We therefore see that the maximum likelihood result of a point estimate for μ is recovered precisely from the Bayesian formalism in the limit of an infinite number of observations.