Parametric Density Estimation:

Maximum Likelihood Estimation

C6

Today

- Introduction to density estimation
  - Maximum Likelihood Estimation

Introducton

- Bayesian Decision Theory in previous lectures tells us how to design an optimal classifier if we knew:
  - $\boldsymbol{P}(\boldsymbol{c}_i)$ (priors)
  - $\boldsymbol{P}(\boldsymbol{x} \mid \boldsymbol{c}_i)$ (class-conditional densities)
- Unfortunately, we rarely have this complete information!

Probability density methods

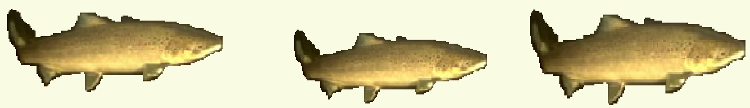
- Parametric methods – assume we know the shape of the distribution, but not the parameters. Two types of parameter estimation:
  - Maximum Likelihood Estimation
  - Bayesian Estimation
- Non parametric methods – the form of the density is entirely determined by the data without any model.

Independence Across Classes

- We have training data for each class

salmon *sea bass* *salmon* *salmon* *sea bass* *sea bass*

- When estimating parameters for one class, will only use the data collected for that class
  - reasonable assumption that data from class $\boldsymbol{c_i}$ gives no information about distribution of class $\boldsymbol{c_j}$

estimate parameters for distribution of salmon from	estimate parameters for distribution of bass from
	

Independence Across Classes

- For each class $\boldsymbol{c_i}$ we have a proposed density $\boldsymbol{p_i}(\boldsymbol{x} | \boldsymbol{c_i})$ with unknown parameters θ^i which we need to estimate
- Since we assumed independence of data across the classes, estimation is an identical procedure for all classes
- To simplify notation, we drop sub-indexes and say that we need to estimate parameters θ for density $\boldsymbol{p}(\boldsymbol{x})$
  - the fact that we need to do so for each class on the training data that came from that class is implied

Maximum Likelihood Parameter Estimation

- Parameters θ are unknown but fixed (i.e. not random variables).
- Given the training data, choose the parameter value θ that makes the data most probable (i.e., maximizes the probability of obtaining the sample that has actually been observed)

Maximum Likelihood Parameter Estimation

- We have density $\boldsymbol{p}(\boldsymbol{x})$ which is completely specified by parameters $\theta = [\theta_1, \ldots, \theta_k]$
  - If $\boldsymbol{p}(\boldsymbol{x})$ is $\boldsymbol{N}(\mu, \sigma^2)$ then $\theta = [\mu, \sigma^2]$
- To highlight that $\boldsymbol{p}(\boldsymbol{x})$ depends on parameters θ we will write $\boldsymbol{p}(\boldsymbol{x}|\theta)$
  - Note overloaded notation, $\boldsymbol{p}(\boldsymbol{x}|\theta)$ is **not** a conditional density
- Let $D=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be the $\boldsymbol{n}$ independent training samples in our data
  - If $\boldsymbol{p}(\boldsymbol{x})$ is $\boldsymbol{N}(\mu, \sigma^2)$ then $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are iid samples from $\boldsymbol{N}(\mu, \sigma^2)$

Maximum Likelihood Parameter Estimation

- Consider the following function, which is called likelihood of θ with respect to the set of samples *D*

$$p(D \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta) = F(\theta)$$

- Maximum likelihood estimate (abbreviated MLE) of θ is the value of θ that maximizes the likelihood function $p(D|\theta)$

$$\boxed{\hat{\theta} = \underset{\theta}{\arg\max}\left(p(D \mid \theta)\right)}$$

ML Parameter Estimation vs. ML Classifier

- Recall ML classifier

fixed data

decide class $\boldsymbol{c_i}$ which maximizes $\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{c_i})$

- Compare with ML parameter estimation

fixed data

choose θ that maximizes $\boldsymbol{p}(\boldsymbol{D}|\theta)$

- ML classifier and ML parameter estimation use the same principles applied to different problems

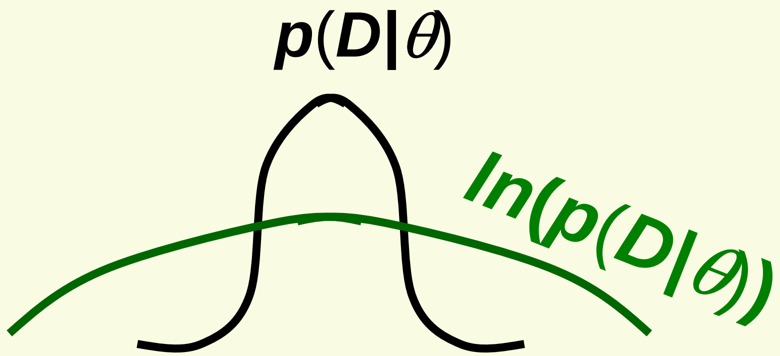
Maximum Likelihood Estimation (MLE)

- Instead of maximizing $p(D|\theta)$, it is usually easier to maximize $\ln(p(D|\theta))$

- Since log is monotonic

$$\hat{\theta} = \arg\max_{\theta}\left(p(D\,|\,\theta)\right) =$$

$$= \arg\max_{\theta}\left(\ln p(D\,|\,\theta)\right)$$



- To simplify notation, $\ln(p(D|\theta))=L(\theta)$

$$\boxed{\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta}\left(\ln\prod_{k=1}^{k=n} p(x_k\,|\,\theta)\right) = \arg\max_{\theta}\left(\sum_{k=1}^{n}\ln p(x_k\,|\,\theta)\right)}$$

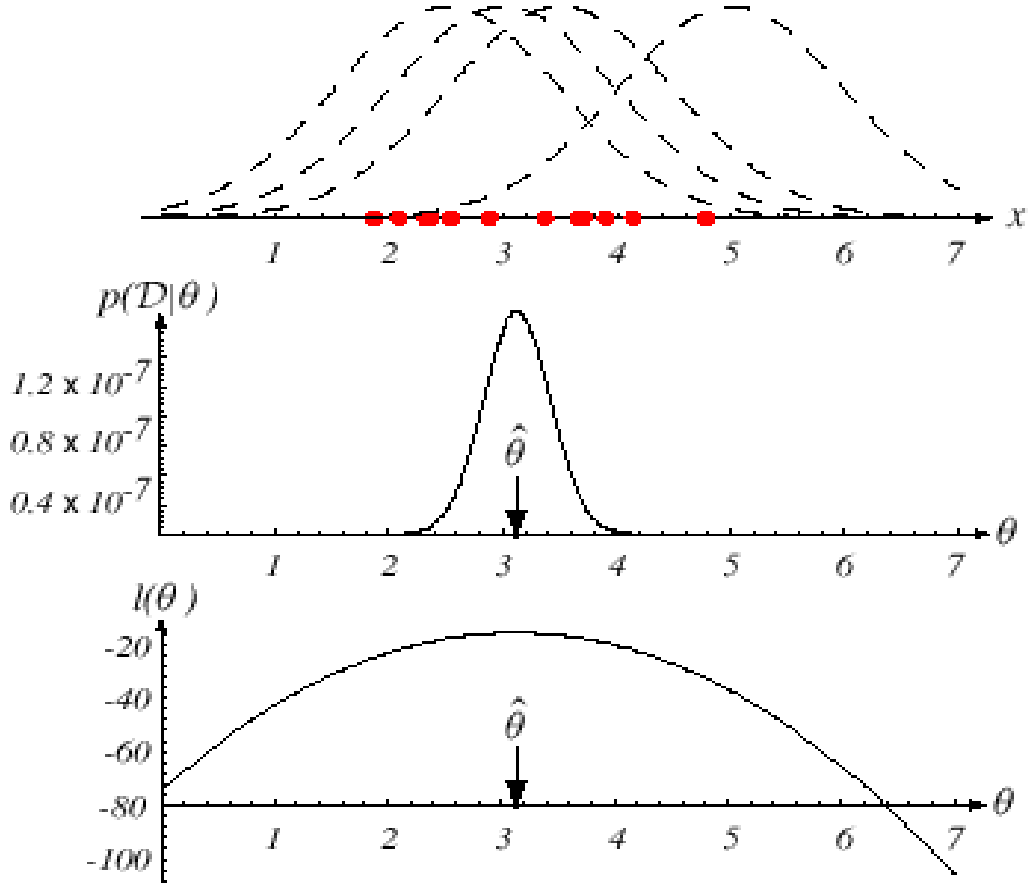


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

MLE: Maximization Methods

- Maximizing $\boldsymbol{L(\theta)}$ can be solved using standard methods from Calculus

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^t$$

- Set of necessary conditions for an optimum is:

$$\nabla_\theta \boldsymbol{l} = \boldsymbol{0}$$

- Also have to check that θ that satisfies the above condition is maximum, not minimum or saddle point. Also check the boundary of range of θ

MLE Example: Gaussian with unknown μ

- Fortunately for us, most of the ML estimates of any densities we would care about have been computed
- Let's go through an example anyway
- Let ***p***(***x***| μ) be ***N***(μ,σ^2) that is σ^2 is known, but μ is unknown and needs to be estimated, so $\theta = \mu$

$$\hat{\mu} = \arg\max_{\mu} L(\mu) = \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln p(x_k \mid \mu)\right) =$$

$$= \arg\max_{\mu}\left(\sum_{k=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right)\right)\right) =$$

$$= \arg\max_{\mu} \sum_{k=1}^{n}\left(-\ln\sqrt{2\pi\sigma} - \frac{(x_k-\mu)^2}{2\sigma^2}\right)$$

MLE Example: Gaussian with unknown μ

$$\arg\max_{\mu}(L(\mu)) = \arg\max_{\mu} \sum_{k=1}^{n} \left(-\ln\sqrt{2\pi\sigma} - \frac{(x_k - \mu)^2}{2\sigma^2} \right)$$

$$\frac{d}{d\mu}(L(\mu)) = \sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) = 0 \quad \Rightarrow \quad \sum_{k=1}^{n} x_k - n\mu = 0 \quad \Rightarrow$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- Thus the ML estimate of the mean is just the average value of the training data, very intuitive!
  - average of the training data would be our guess for the mean even if we didn't know about ML estimates

MLE for Gaussian with unknown μ, σ^2

- Similarly it can be shown that if $\boldsymbol{p}(\boldsymbol{x}|\ \mu,\sigma^2)$ is $\boldsymbol{N}(\mu, \sigma^2)$, that is x both mean and variance are unknown, then again very intuitive result

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

- Similarly it can be shown that if $\boldsymbol{p}(\boldsymbol{x}|\ \mu,\Sigma)$ is $\boldsymbol{N}(\mu, \Sigma)$, that is $\boldsymbol{x}$ is a multivariate Gaussian with both mean and covariance matrix unknown, then

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

How to Measure Performance of MLE?

- How good is a ML estimate $\hat{\theta}$?
  - or actually any other estimate of a parameter?
- The natural measure of error would be $\left|\theta - \hat{\theta}\right|$
- But $\left|\theta - \hat{\theta}\right|$ is random, we cannot compute it before we carry out experiments
  - We want to say something meaningful about our estimate as a function of θ
- A way to solve this difficulty is to ***average*** the error, i.e. compute the ***mean absolute error***

$$E[|\theta - \hat{\theta}|] = \int |\theta - \hat{\theta}|\, p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n$$

How to Measure Performance of MLE?s

- It is usually much easier to compute an almost equivalent measure of performance, the ***mean squared error***:

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

- Do a little algebra, and use Var(***X***)=E(***X²***)-(E(***X***))²

$$E\left[\left(\theta-\hat{\theta}\right)^2\right]=\underbrace{Var\left(\hat{\theta}\right)}_{\text{variance}}+\underbrace{\left(E\left(\hat{\theta}\right)-\theta\right)^2}_{\text{bias}}$$

variance

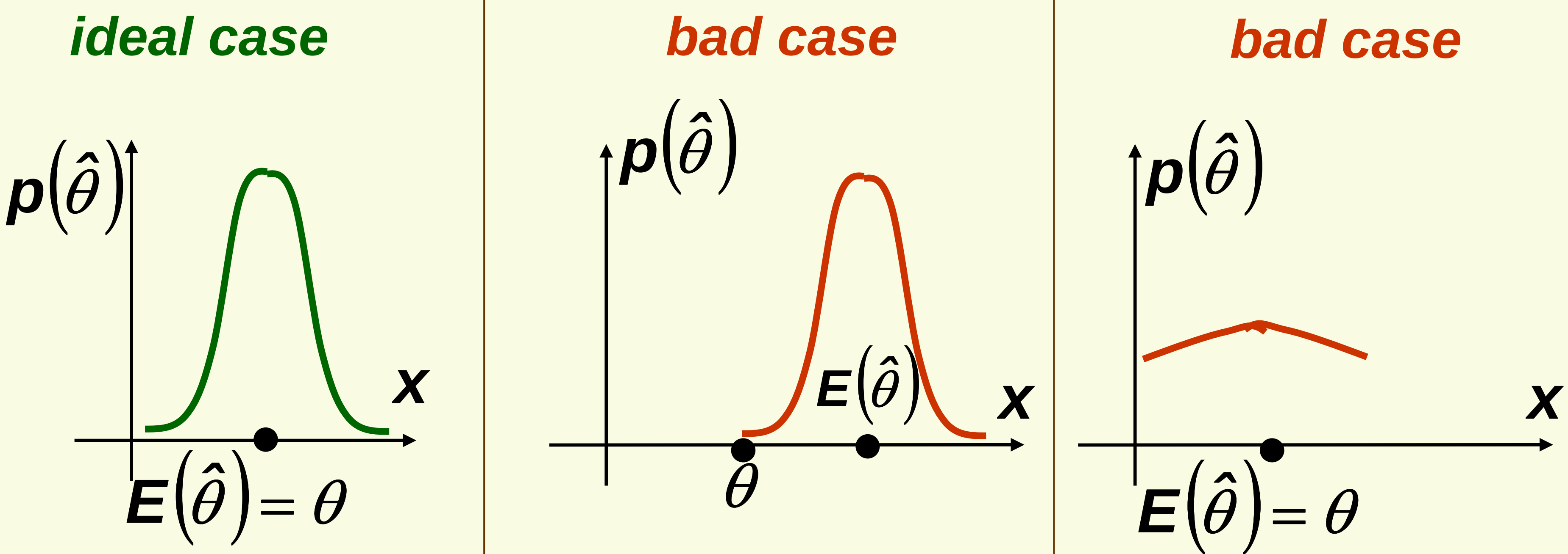
estimator should have low variance

bias

expectation should be close to the true θ

How to Measure Performance of MLE?

$$E\left[\left(\theta-\hat{\theta}\right)^2\right] = \underbrace{Var\left(\hat{\theta}\right)}_{\textbf{variance}} + \underbrace{\left(E\left(\hat{\theta}\right)-\theta\right)^2}_{\textbf{bias}}$$



no bias
low variance

large bias
low variance

no bias
high variance

Bias and Variance for MLE of the Mean

- Let's compute the bias for ML estimate of the mean

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{k=1}^{n} x_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[x_k] = \frac{1}{n}\sum_{k=1}^{n}\mu = \mu$$

  - Thus this estimate is unbiased!

- How about variance of ML estimate of the mean?

$$E\left[(\hat{\mu}-\mu)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right]^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\mu)(x_j-\mu)\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[(x_i-\mu)(x_j-\mu)\right]$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- Thus variance is very small for a large number of samples (the more samples, the smaller is variance)

- Thus the MLE of the mean is a very good estimator

Bias and Variance for MLE of the Mean

- Suppose someone claims they have a new great estimator for the mean, just take the first sample!

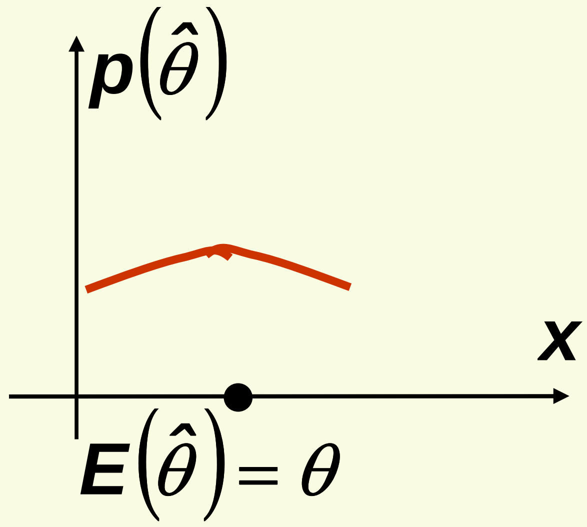
$$\hat{\mu} = x_1$$

- Thus this estimator is unbiased: $E(\hat{\mu}) = E(x_1) = \mu$

- However its variance is:

$$E\left[(\hat{\mu} - \mu)^2\right] = E\left[(x_1 - \mu)^2\right] = \sigma^2$$

- Thus variance can be very large and does not improve as we increase the number of samples



no bias
high variance

MLE Bias for Mean and Variance

- How about ML estimate for the variance?

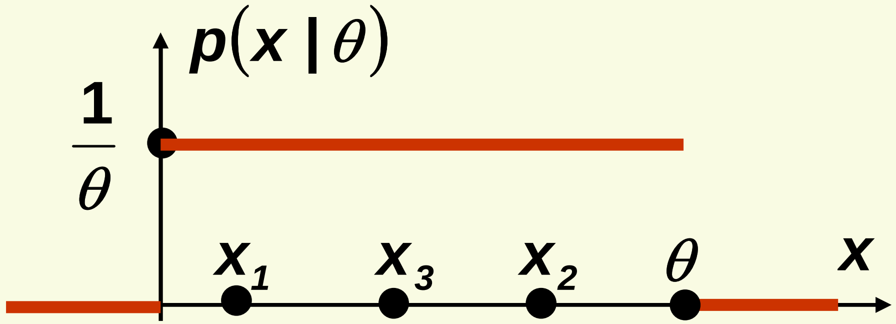
$$E\left[\hat{\sigma}^2\right] = E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

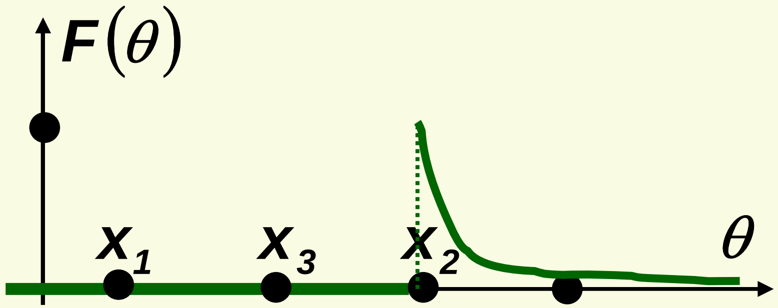
  - Thus this estimate is biased!
    - This is because we used $\hat{\mu}$ instead of true μ
  - Bias →0 as n→ infinity, *asymptotically* unbiased
  - Unbiased estimate $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$

- Variance of MLE of variance can be shown to go to 0 as n goes to infinity

MLE for Uniform distribution U[0,θ]

- X is U[0,θ] if its density is 1/θ inside [**0**,θ] and 0 otherwise (uniform distribution on [**0**,θ])





- The likelihood is $F(\theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq \max\{x_1, \ldots, x_n\} \\ 0 & \text{if } \theta < \max\{x_1, \ldots, x_n\} \end{cases}$

- Thus $\hat{\theta} = \arg\max_{\theta} \left(\prod_{k=1}^{k=n} p(x_k \mid \theta) \right) = \max\{x_1, \ldots, x_n\}$

- This is not very pleasing since for sure θ should be larger than any observed ***x***!