CS434a/541a: Pattern Recognition
Prof. Olga Veksler

Lecture 5
Today

- Introduction to parameter estimation
- Two methods for parameter estimation
  - Maximum Likelihood Estimation
  - Bayesian Estimation
Bayesian Decision Theory in previous lectures tells us how to design an optimal classifier if we knew:

- $P(c_i)$ (priors)
- $P(x | c_i)$ (class-conditional densities)

Unfortunately, we rarely have this complete information!

Suppose we know the shape of distribution, but not the parameters

- Two types of parameter estimation
  - Maximum Likelihood Estimation
  - Bayesian Estimation
**ML and Bayesian Parameter Estimation**

- Shape of probability distribution is known
  - Happens sometimes
- Labeled training data
- Need to estimate parameters of probability distribution from the training data

---

**Example**

respected fish expert says salmon’s length has distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ and sea bass’s length has distribution $\mathcal{N}(\mu_2, \sigma_2^2)$

- Need to estimate parameters $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$
- Then design classifiers according to the bayesian decision theory
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 $c_i$ gives no information about distribution of class $c_j$

estimate parameters for distribution of salmon from

estimate parameters for distribution of bass from
Independence Across Classes

- For each class $c_i$ we have a proposed density $p_i(x/ c_i)$ with unknown parameters $\theta^i$ 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 $\theta$ for density $p(x)$.
  - the fact that we need to do so for each class on the training data that came from that class is implied.
ML vs. Bayesian Parameter Estimation

- **Maximum Likelihood**
  - Parameters $\theta$ are unknown but fixed (i.e. not random variables)

- **Bayesian Estimation**
  - Parameters $\theta$ are random variables having some known a priori distribution (prior)

- After parameters are estimated with either ML or Bayesian Estimation we use methods from Bayesian decision theory for classification
We have density $p(x)$ which is completely specified by parameters $\theta = [\theta_1, \ldots, \theta_k]$
- If $p(x)$ is $N(\mu, \sigma^2)$ then $\theta = [\mu, \sigma^2]$

To highlight that $p(x)$ depends on parameters $\theta$ we will write $p(x|\theta)$
- Note overloaded notation, $p(x|\theta)$ is not a conditional density

Let $D = \{x_1, x_2, \ldots, x_n\}$ be the $n$ independent training samples in our data
- If $p(x)$ is $N(\mu, \sigma^2)$ then $x_1, x_2, \ldots, x_n$ are iid samples from $N(\mu, \sigma^2)$
Consider the following function, which is called likelihood of $\theta$ with respect to the set of samples $D$

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

Note if $D$ is fixed $p(D|\theta)$ is not a density.

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

$$\hat{\theta} = \arg\max_{\theta}(p(D | \theta))$$
**Maximum Likelihood Estimation (MLE)**

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

- If \( D \) is allowed to vary and \( \theta \) is fixed, by independence \( p(D \mid \theta) \) is the joint density for \( D=\{x_1, x_2, \ldots, x_n\} \)

- If \( \theta \) is allowed to vary and \( D \) is fixed, \( p(D \mid \theta) \) is not density, it is likelihood \( F(\theta)! \)

- Recall our approximation of integral trick

\[
Pr[D \in B[x_1, \ldots, x_n] \mid \theta] \approx \varepsilon \prod_{k=1}^{k=n} p(x_k \mid \theta)
\]

- Thus ML chooses \( \theta \) that is most likely to have given the observed data \( D \)
ML Parameter Estimation vs. ML Classifier

- Recall ML classifier
  decide class $c_i$ which maximizes $p(x|c_i)$

- Compare with ML parameter estimation
  choose $\theta$ that maximizes $p(D|\theta)$

- 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 minimize $\ln(p(D|\theta))$

- Since log is monotonic
  \[
  \hat{\theta} = \arg \max_{\theta} (p(D | \theta)) = \\
  = \arg \max_{\theta} (\ln p(D | \theta))
  \]

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

\[
\hat{\theta} = \arg \max_{\theta} l(\theta) = \arg \max_{\theta} \left( \ln \prod_{k=1}^{k=n} p(x_k | \theta) \right) = \arg \max_{\theta} \left( \sum_{k=1}^{n} \ln p(x_k | \theta) \right)
\]
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 $p(D|\theta)$ 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 $p(D|\theta)$ is shown as a function of $\theta$ whereas the conditional density $p(x|\theta)$ is shown as a function of $x$. Furthermore, as a function of $\theta$, the likelihood $p(D|\theta)$ 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*. Copyright © 2001 by John Wiley & Sons, Inc.
**MLE: Maximization Methods**

- Maximizing $l(\theta)$ can be solved using standard methods from Calculus

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let $\nabla_\theta$ 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:

  $$\nabla_\theta l = 0$$

- Also have to check that $\theta$ that satisfies the above condition is maximum, not minimum or saddle point. Also check the boundary of range of $\theta$
MLE Example: Gaussian with unknown $\mu$

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

$$
\hat{\mu} = \arg \max_{\mu} l(\mu) = \arg \max_{\mu} \left( \sum_{k=1}^{n} \ln p(x_k | \mu) \right) = \\
= \arg \max_{\mu} \left( \sum_{k=1}^{n} \ln \left( \frac{1}{\sqrt{2\pi}\sigma} \exp \left( - \frac{(x_k - \mu)^2}{2\sigma^2} \right) \right) \right) = \\
= \arg \max_{\mu} \sum_{k=1}^{n} \left( -\ln \sqrt{2\pi}\sigma - \frac{(x_k - \mu)^2}{2\sigma^2} \right)
$$
MLE Example: Gaussian with unknown $\mu$

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

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

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

- 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 $\mu, \sigma^2$**

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

\[
\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})^2
\]

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

\[
\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k \quad \hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^t
\]
Today

- Finish Maximum Likelihood Estimation
- Bayesian Parameter estimation
- New Topic
  - Non Parametric Density Estimation
How good is a ML estimate $\hat{\theta}$? or actually any other estimate of a parameter?

The natural measure of error would be $|\theta - \hat{\theta}|$

But $|\theta - \hat{\theta}|$ is random, we cannot compute it before we carry out experiments.

We want to say something meaningful about our estimate as a function of $\theta$.

A way to solve this difficulty is to average the error, i.e. compute the **mean absolute error**

$$E|\theta - \hat{\theta}| = \int |\theta - \hat{\theta}| p(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n$$
How to Measure Performance of MLEs

- It is usually much easier to compute an almost equivalent measure of performance, the mean squared error:
  \[ E[(\theta - \hat{\theta})^2] \]

- Do a little algebra, and use
  \[ \text{Var}(X) = E(X^2) - (E(X))^2 \]
  \[ E[(\theta - \hat{\theta})^2] = \text{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \]

  - variance: estimator should have low variance
  - bias: expectation should be close to the true \( \theta \)
How to Measure Performance of MLE's

\[ E[(\theta - \hat{\theta})^2] = \text{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \]

- **ideal case**
  - \( p(\hat{\theta}) \)
  - \( E(\hat{\theta}) = \theta \)
  - **no bias**
  - **low variance**

- **bad case**
  - \( p(\hat{\theta}) \)
  - **large bias**
  - **low variance**

- **bad case**
  - \( p(\hat{\theta}) \)
  - \( E(\hat{\theta}) = \theta \)
  - **no bias**
  - **high variance**
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?

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

\[ = \frac{\sigma^2}{n} \]

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

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

\[ \text{no bias high variance} \]
MLE Bias for Mean and Variance

- How about ML estimate for the variance?

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

- Thus this estimate is biased!
  - This is because we used \( \hat{\mu} \) instead of true \( \mu \)
  - Bias \( \to 0 \) as \( n \to \infty \), 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])

```
\[
p(x / θ) = \begin{cases} 
  \frac{1}{θ} & \text{if } θ \geq \max\{x_1, ..., x_n\} \\
  0 & \text{if } θ < \max\{x_1, ..., x_n\}
\end{cases}
\]
```

- The likelihood is

```
\[
F(θ) = \prod_{k=1}^{k=n} p(x_k | θ) = \begin{cases} 
  \frac{1}{θ^n} & \text{if } θ \geq \max\{x_1, ..., x_n\} \\
  0 & \text{if } θ < \max\{x_1, ..., x_n\}
\end{cases}
\]
```

- Thus

```
\[
\hat{θ} = \arg \max_θ \left( \prod_{k=1}^{k=n} p(x_k | θ) \right) = \max\{x_1, ..., x_n\}
\]
```

- This is not very pleasing since for sure θ should be larger than any observed x!
Suppose we have some idea of the range where parameters $\theta$ should be

- Shouldn’t we formalize such prior knowledge in hopes that it will lead to better parameter estimation?

- Let $\theta$ be a random variable with prior distribution $P(\theta)$
  - This is the key difference between ML and Bayesian parameter estimation
  - This key assumption allows us to fully exploit the information provided by the data
Bayesian Parameter Estimation

- As in MLE, suppose $p(x|\theta)$ is completely specified if $\theta$ is given.
- But now $\theta$ is a random variable with prior $p(\theta)$
  - Unlike MLE case, $p(x|\theta)$ is a conditional density.
- After we observe the data $D$, using Bayes rule we can compute the posterior $p(\theta|D)$.
- Recall that for the MAP classifier we find the class $c_i$ that maximizes the posterior $p(c|D)$.
- By analogy, a reasonable estimate of $\theta$ is the one that maximizes the posterior $p(\theta|D)$.
- But $\theta$ is not our final goal, our final goal is the unknown $p(x)$.
- Therefore a better thing to do is to maximize $p(x|D)$, this is as close as we can come to the unknown $p(x)$!
**Bayesian Estimation: Formula for** \( p(x|D) \)

- From the definition of joint distribution:

\[
p(x | D) = \int p(x, \theta | D) d\theta
\]

- Using the definition of conditional probability:

\[
p(x | D) = \int p(x | \theta, D)p(\theta | D) d\theta
\]

- But \( p(x|\theta,D)=p(x|\theta) \) since \( p(x|\theta) \) is completely specified by \( \theta \)

\[
p(x | D) = \int \begin{array}{c} \text{known} \\ p(x | \theta) \end{array} \begin{array}{c} \text{unknown} \\ p(\theta | D) \end{array} d\theta
\]

- Using Bayes formula,

\[
p(\theta | D) = \frac{p(D | \theta)p(\theta)}{\int p(D | \theta)p(\theta) d\theta}
\]

\[
p(D | \theta) = \prod_{k=1}^{n} p(x_k | \theta)
\]
Bayesian Estimation vs. MLE

- So in principle $p(x|D)$ can be computed
  - In practice, it may be hard to do integration analytically, may have to resort to numerical methods

$$p(x / D) = \int p(x / \theta) \frac{\prod_{k=1}^{n} p(x_k / \theta)p(\theta)}{\int \prod_{k=1}^{n} p(x_k / \theta)p(\theta)d\theta} \, d\theta$$

- Contrast this with the MLE solution which requires differentiation of likelihood to get $p(x / \theta)$
  - Differentiation is easy and can always be done analytically
**Bayesian Estimation vs. MLE**

- $p(x|D)$ can be thought of as the weighted average of the proposed model all possible values of $\theta$ with certain support $\theta$ receives from the data.

\[
p(x | D) = \int p(x | \theta) p(\theta | D) d\theta
\]

- Contrast this with the MLE solution which always gives us a single model:

\[
p(x / \hat{\theta})
\]

- When we have many possible solutions, taking their sum averaged by their probabilities seems better than spitting out one solution.
Bayesian Estimation: Example for $U[0, \theta]$

- Let $X$ be $U[0, \theta]$. Recall $p(x|\theta) = 1/\theta$ inside $[0, \theta]$, else 0

- Suppose we assume a $U[0,10]$ prior on $\theta$
  - good prior to use if we just know the range of $\theta$ but don’t know anything else

- We need to compute $p(x|D) = \int p(x|\theta)p(\theta|D)d\theta$
  - with $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$ and $p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$
Bayesian Estimation: Example for U[0, \theta]

- We need to compute \( p(x \mid D) = \int p(x \mid \theta) p(\theta \mid D) d\theta \)

- using \( p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta) d\theta} \) and \( p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta) \)

- When computing MLE of \( \theta \), we had
  \[
p(D \mid \theta) = \begin{cases} 
    \frac{1}{\theta^n} & \text{for } \theta \geq \max\{x_1, \ldots, x_n\} \\
    0 & \text{otherwise}
  \end{cases}
\]

- Thus
  \[
p(\theta \mid D) = \begin{cases} 
    c \frac{1}{\theta^n} & \text{for } \max\{x_1, \ldots, x_n\} \leq \theta \leq 10 \\
    0 & \text{otherwise}
  \end{cases}
\]

- where \( c \) is the normalizing constant, i.e.
  \[
c = \frac{1}{\int_{\max\{x_1, \ldots, x_n\}}^{10} \frac{d\theta}{\theta^n}}
\]
Bayesian Estimation: Example for U[0, \theta]

- We need to compute
  \[ p(x \mid D) = \int p(x \mid \theta)p(\theta \mid D) d\theta \]
  \[ p(\theta \mid D) = \begin{cases} 
    c \frac{1}{\theta^n} & \text{for } \max\{x_1, \ldots, x_n\} \leq \theta \leq 10 \\
    0 & \text{otherwise}
  \end{cases} \]

- We have 2 cases:
  1. case \( x < \max\{x_1, x_2, \ldots, x_n\} \)
     \[ p(x \mid D) = \int_{\max\{x_1, \ldots, x_n\}}^{10} c \frac{1}{\theta^{n+1}} d\theta = \alpha \]
  2. case \( x > \max\{x_1, x_2, \ldots, x_n\} \)
     \[ p(x \mid D) = \int_{x}^{10} c \frac{1}{\theta^{n+1}} d\theta = \left. \frac{c}{-n\theta^n} \right|_{x}^{10} = \frac{c}{nx^n} - \frac{c}{n10^n} \]
- Note that even after $x > \max \{x_1, x_2, \ldots, x_n\}$, Bayes density is not zero, which makes sense.
- curious fact: Bayes density is not uniform, i.e. does not have the functional form that we have assumed!
ML vs. Bayesian Estimation with Broad Prior

- Suppose \( p(\theta) \) is flat and broad (close to uniform prior)
- \( p(\theta|D) \) tends to sharpen if there is a lot of data
- Thus \( p(D|\theta) \propto p(\theta|D)/p(\theta) \) will have the same sharp peak as \( p(\theta|D) \)
- But by definition, peak of \( p(D|\theta) \) is the ML estimate \( \hat{\theta} \)
- The integral is dominated by the peak:
  \[
p(x / D) = \int p(x / \theta)p(\theta / D)d\theta \approx p(x / \hat{\theta})\int p(\theta / D)d\theta = p(x / \hat{\theta})
\]
- Thus as \( n \) goes to infinity, Bayesian estimate will approach the density corresponding to the MLE!
ML vs. Bayesian Estimation: General Prior

- **Maximum Likelihood Estimation**
  - Easy to compute, use differential calculus
  - Easy to interpret (returns one model)
  - \( p(\hat{x}/\theta) \) has the assumed parametric form

- **Bayesian Estimation**
  - Hard to compute, need multidimensional integration
  - Hard to interpret, returns weighted average of models
  - \( p(x/D) \) does not necessarily have the assumed parametric form
  - Can give better results since use more information about the problem (prior information)