Note

- Homework #3 Due now.
- Project topic/papers due next week.
  - Read the assignment thoroughly
  - Submit to iLearn the topic choice and selected papers for
    my review. **Late policy will apply.**
- Fast Prototyping Exercise #2 on Mean Shift
  starts today.
  - [https://bidal.sfsu.edu/~kazokada/csc872/PD2.pdf](https://bidal.sfsu.edu/~kazokada/csc872/PD2.pdf)
  - [https://bidal.sfsu.edu/~kazokada/csc872/Segmentation_Data.zip](https://bidal.sfsu.edu/~kazokada/csc872/Segmentation_Data.zip)

---

**Parametric Statistical Modeling**

CSC 872
Pattern Analysis and Machine Intelligence

References
Andrew Moore’s great slides at
http://www.cs.cmu.edu/~awm/tutorials
PF: Statistical Modeling: Review

• Problem
  – Estimating probability distribution from data
  – Data = Samples drawn from an unknown underlying distribution
  – What is the underlying distribution given these data?

\[ P(X) \]
\[ \{x_1, \ldots, x_n\} \]
\[ X \]

Non-Parametric Modeling: Review

• Histogram & Kernel Density Estimation
  – No prior assumption about the density function

• Advantages
  – Flexible (for any shape of distribution)

• Disadvantages
  – Needs Quantization or Bandwidth Parameter Tuning
  – High Time and Space Complexity
  – Needs to store all data points for KDE!
  – Takes a lot of time to build and use these things
New Strategy: Parametric Modeling

• Let’s use prior knowledge/assumption of the target distribution !!!

• Two-Step Strategy
  • (1) Choose A Parameterized Function
    – Pick a function with parameters that control its shape and location
    – It is up to us what function we use
    – You need to choose the function according to your prior knowledge!!!
  • (2) Do Parameter Estimation
    – Basically fit the function to the data ...in another words ...
    – Estimate the parameters that make the function fit best to the data

• Why we do this?
  • Parameters are typically much fewer so...
    (1) Greatly improve time and space complexity
  • Parameter Estimation is a well-studied field
    (2) Nice mathematical framework that is called...
PF: Maximum Likelihood Estimation

- **Maximum Likelihood Estimate (MLE)**

- You saw this first in Bayesian Reasoning Lec
- **Foundation** of pattern analysis and learning
- NOT Bayesian Inference!
- **Maximum A Posteriori Estimate (MAP)**
- MLE is used more than MAP
- Why?
- I get back to this later

---

PF: Maximum Likelihood Estimation !!!

- **Suppose we have independent and identically distributed samples drawn from a distribution parameterized by** $\alpha$
  - $x_1, x_2, \ldots, x_N \sim \text{(i.i.d.) } p(x|\alpha) := f(x, \alpha)$
  - You know a form of $f$ BUT you don’t know the value of $\alpha$

- **For what $\alpha$ are these samples most likely?**

$$\alpha^{mle} = \arg\max_{\alpha} p(x_1, \ldots, x_N | \alpha)$$
MLE for Gaussian (Normal) Mean

• Suppose we have $x_1, \ldots, x_N \sim (\text{i.i.d.}) N(\mu, \sigma^2)$
• But you don’t know $\mu$ (known $N$ and $\sigma^2$)
• MLE: For which $\mu$ is $x_1, \ldots, x_N$ most likely?

$$\mu_{\text{MLE}} = \arg\max_{\mu} p(x_1, \ldots, x_N | \mu, \sigma^2)$$

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Algebra & Calculus to simplify the problem & Solving $\nabla p(x) = 0$ to maximize the likelihood

Some Algebra

$\mathcal{X}_1 \perp \mathcal{X}_2 \implies p(x_1, x_2) = p(x_1) p(x_2)$

$$\mu_{\text{MLE}} = \arg\max_{\mu} p(x_1, \ldots, x_N | \mu, \sigma^2) \quad \text{MLE}$$

$$= \arg\max_{\mu} \prod_{n=1}^{N} p(x_n | \mu, \sigma^2) \quad \text{i.i.d. assumption}$$

$$= \arg\max_{\mu} \log \left[ \prod_{n=1}^{N} p(x_n | \mu, \sigma^2) \right] \quad \text{Log monotonisity}$$

$$= \arg\max_{\mu} \sum_{n=1}^{N} \log[p(x_n | \mu, \sigma^2)]$$

Log-likelihood !!!

$$= \arg\max_{\mu} \sum_{n=1}^{N} -\frac{(x_n-\mu)^2}{2\sigma^2} + C$$

Plugging in Gaussian

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \arg\min_{\mu} \sum_{n=1}^{N} (x_n - \mu)^2$$

Removing parts that are not related to the optimization
More Algebra

\[ \mu^{mle} = \underset{\mu}{\text{argmax}} p(x_1, \ldots, x_N | \mu, \sigma^2) \]

\[ = \arg\min_{\mu} \sum_{n=1}^{N} (x_n - \mu)^2 \]

Solving \[ \frac{\partial}{\partial \mu} \left[ \sum_{n=1}^{N} (x_n - \mu)^2 \right] \]

Argmin/Argmax is

\[ 0 = \sum_{n=1}^{N} 2(x_n - \mu^{mle}) \]

Do differentiation

\[ \mu^{mle} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

Solve it about \( \mu \)

What do we get?

- MLE \( \mu^{mle} \) of a normal distribution is a **sample mean**

\[ \mu^{mle} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

- In another word
- Computing the sample mean =
- Computing MLE of the true population mean =
- Computing MLE of the center of a Gaussian fitted to your data
How do we do that? (Recipe for MLE)

- **TASK:** Find $\theta$ assuming known form of $p(\text{Data}|\theta,...)$

1. Derive **log-likelihood** (LL): $LL = \log p(\text{Data}|\theta,...)$
2. Do **calculus/algebra** on $\partial LL/\partial \theta$
3. Create an equation by **setting** $\partial LL/\partial \theta = 0$
4. **Solve** $\partial LL/\partial \theta = 0$ about $\theta$ for maximizing $p(\text{Data}|\theta,...)$
5. **Check if the solution is a maximum** instead of minimum or saddle point

For more than one parameters

- **TASK:** Find $\theta$ assuming known form of $p(\text{Data}|\theta=\theta_1,...,\theta_n)$

1. Derive **log-likelihood** (LL): $LL = \log p(\text{Data}|\theta_1,...,\theta_n)$
2. Do **calculus/algebra** on $\partial LL/\partial \theta_1,\ldots,\partial LL/\partial \theta_n$
3. Create a set of equations by setting
   \[
   \begin{align*}
   \partial LL/\partial \theta_1 &= 0 \\
   \partial LL/\partial \theta_2 &= 0 \\
   \vdots \\
   \partial LL/\partial \theta_n &= 0
   \end{align*}
   \]
4. **Solve the simultaneous equations** about $\theta=\theta_1,...,\theta_n$
5. **Check if the solution is a maximum**
Back to MLE of Gaussian

$$\theta^{mle} = \arg \max_{\theta} p(x_1, \ldots, x_N | \theta = (\mu, \sigma^2))$$

$$LL = \log[p(x_1, \ldots, x_N | \mu, \sigma^2)]$$

$$=-0.5N\log2\pi - 0.5N\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\frac{\partial L}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0 \quad \Rightarrow \quad \mu^{mle} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

$$\frac{\partial LL}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 = 0 \quad \Rightarrow \quad \sigma^2_{mle} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu^{mle})^2$$

Sample variance!!!

Unbiased Estimator

• Unbiased Estimator
  – **Expected value** of the estimate is the same as the **true value** of the estimate

• Suppose $x_1, \ldots, x_N \sim$ (i.i.d.) $N(\mu, \sigma^2)$

$$\mu^{mle} = \frac{1}{N}\sum_{n=1}^{N}x_n \quad E[x] = \mu$$

$$E[\mu^{mle}] = E\left[\frac{1}{N}\sum_{n=1}^{N}x_n\right] = \frac{1}{N}\sum_{n=1}^{N}E[x_n] = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu$$

$$E[\mu^{mle}] = \mu \quad \mu^{mle} \text{ is unbiased}$$
Biased Estimator

• Biased Estimator
  – **Expected value** of the estimate is **different from** the **true value** of the estimate

• Suppose \( x_1, \ldots, x_N \sim \text{(i.i.d.) } N(\mu, \sigma^2) \)

\[
\sigma_{\text{mle}}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{mle}})^2
\]

\[
E[\sigma_{\text{mle}}^2] = E\left[ \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{mle}})^2 \right] = \frac{1}{N} E \left[ \sum x_n^2 - \frac{1}{N} \sum_{n} \sum_{n'} x_n x_{n'} \right]
\]

\[
= \frac{1}{N^2} E \left[ (N-1) \sum x_n^2 - \sum_{n \neq n'} x_n x_{n'} \right]
\]

\[
= \frac{N-1}{N^2} \sum n E[x_n^2] - N E[x_n]^2 = \frac{N-1}{N} \sigma^2
\]

\[E[x^2] - E[x]^2 = \sigma^2\]

\[\sigma_{\text{mle}}^2 \neq \sigma^2\]

\(\sigma_{\text{mle}}^2\) is biased!!!
What we gonna do?

• Bias: $E[\theta^{mle}] - \theta$
  
  $Bias[\sigma^2_{mle}] = E[\sigma^2_{mle}] - \sigma^2 = \frac{N-1}{N} \sigma^2 - \sigma^2 = -\frac{1}{N} \sigma^2$

• Unbiased estimator from a biased one
  
  $E[\sigma^2_{mle}] = \frac{N-1}{N} \sigma^2$  \quad  $E[\sigma^2_{unbiased}] = \sigma^2$

  $\sigma^2_{unbiased} = \frac{N}{N-1} \sigma^2_{mle} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu)^2$

• Is unbiased estimator always better?

  $\mu^{est1} = x_4$

  $\mu^{est2} = \frac{1}{N + 10} \sum_{n=1}^{N} x_n$

• Asymptotically unbiased estimator
  
  $E[\sigma^2_{mle}] = \frac{N-1}{N} \sigma^2 \rightarrow \sigma^2$

And more … what if

• TASK: Find $\theta$ assuming known form of $p(Data|\theta_1,...,\theta_n)$

  1. Derive log-likelihood (LL): $LL = \log p(Data|\theta_1,...,\theta_n)$
  2. Do calculus/algebra on $\partial LL/\partial \theta_1, ..., \partial LL/\partial \theta_n$
  3. Create an equation by setting

     \[ \begin{align*}
      \frac{\partial LL}{\partial \theta_1} &= 0 \\
      \frac{\partial LL}{\partial \theta_2} &= 0 \\
      \vdots \\
      \frac{\partial LL}{\partial \theta_n} &= 0 \\
    \end{align*} \]

     What if we cannot solve them???

  4. Solve the simultaneous equations about $\theta=\theta_1,...,\theta_n$
  5. Check if the solution is a maximum
Alternative to Our MLE Recipe

- **Bad News:** for many functions you choose, you **CANNOT SOLVE** the simultaneous equations \( \frac{\partial LL}{\partial \theta_1} = 0, \ldots, \frac{\partial LL}{\partial \theta_n} = 0 \)
- Oh no …
- But there is a savior …
- **Go Iterative !!!**
  (Examples: Mean Shift, EM Algorithm)
- **Variational Method (below is the recipe)**
  - Define a simplified problem using inequality … in another words …
  - Define an analytical lower-bound of your complex density function
  - Find MLE of the (quadratic) lower-bound
  - This solution provides an iterative step (like mean shift vector!)
  - A sequence of this iterator can be proven to asymptotically converge to a nearest mode of the density function

Maximum A Posteriori Estimation

- Suppose we have \( x_1, \ldots, x_N \sim \text{(i.i.d.) } p(x|\theta) \)
- But you don’t know \( \theta \)
- **MLE:** For which \( \theta \) is \( x_1, \ldots, x_N \) most likely?
- **MAP:** Which \( \theta \) maximizes posterior \( p(\theta | x_1, \ldots, x_N) \)

\[
\theta^{mle} = \arg\max_{\theta} p(x_1, \ldots, x_N | \theta) \\
\theta^{map} = \arg\max_{\theta} p(\theta | x_1, \ldots, x_N) \\
= \arg\max_{\theta} \frac{p(x_1, \ldots, x_N | \theta) p(\theta)}{\int_{\theta} p(x_1, \ldots, x_N | \theta) p(\theta) \, d\theta} \\
= \arg\max_{\theta} p(x_1, \ldots, x_N | \theta) p(\theta) \\
\text{Baye's rule}
\]

You need to provide a prior distribution.
MAP for Gaussian (Normal) Mean

- Suppose we have $x_1, \ldots, x_N \sim (i.i.d.) \mathcal{N}(\mu, \sigma^2)$
- But you don’t know $\mu$
- MAP: Which $\mu$ maximizes posterior $p(\mu | x_1, \ldots, x_N, \sigma^2)$
- Set the prior also as a Gaussian $\mathcal{N}(\mu_0, \sigma_0^2)$
  \[
  \mu^{map} = \operatorname{argmax}_\mu p(\mu | x_1, \ldots, x_N, \sigma^2)
  = \operatorname{argmax}_\mu p(x_1, \ldots, x_N | \mu, \sigma^2) p(\mu)
  = \operatorname{argmax}_\mu \mathcal{N}(\mu; \mu_0, \sigma_0^2) \prod_{n=1}^{N} \mathcal{N}(x_n; \mu, \sigma^2)
  = \operatorname{argmax}_\mu \mathcal{N}(\mu; \mu_1, \sigma_1^2)
  = \mu_1 = \frac{\sigma^2 \mu_0 + \sigma_0^2 \sum_{n=1}^{N} x_n}{\sigma^2 + \sigma_0^2 N}
  \]
  \[
  \sigma_1^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + \sigma_0^2 N}
  \]

**Great! But why not MAP?**

- Why did we choose Gaussian as a prior?
- Well, we did not need to really … but
- Because of its analytical simplicity, meaning
- You get a posterior as the same form as prior!!
- **Conjugate Prior**
- This allows us use our MLE recipe to get a closed-form soln.
- For more complex distributions, algebra gets bad (your headache)
- So why not MAP
  - Too much algebra being too much (simpler better!)
  - But really, for larger $N$, it may not differ much from MLE!
  - But really, my nice conjugate prior does not represent my specific problem
  - But really, for my specific problem, I don’t have conjugate prior
PF: Is it learning?

• Just some algebra to derive formulae?
  – Yes, but at the end, you really get a probability distribution approximated by your function whose shape is fit to your data.
  – So these provides a valid means for probabilistic learning!
  – if you are lucky and you get a closed-form solution
  – If not, go ITERATIVE NUMERICAL! (e.g., mean shift / EM)

• What designer must choose?
  – Function form of distributions (likelihood (+ prior))
  – Type of estimation (MLE or MAP or Iterative)

• What must be derived from data?
  – Parameter values
  – Approximated Distributions in an Analytical Form (if you are lucky)

Summary

• Parametric Statistical Modeling
  – Statistical Modeling via Parameter Estimation
  – MLE: Maximum Likelihood Estimation
  – MAP: Maximum A Posteriori Estimation
  – Gaussian is your friend! (or analytical nature of your function matters)
  – For more complex distributions, you can go iterative.
  – DISADVANTAGE:
    – How to pick right function to your data?
    – What if my data does not fit the function I want to choose (e.g., Gaussian)?
    – Most useful function like Gaussian has a limited expression power…

• Next
  – Mixture Model: Parametric Clustering & EM-Algorithm
  – Pattern Classification: PCA and LDA
Mixture of Gaussian

Expectation Maximization Algorithm
Relation to Clustering