Note

$p(h|d)$ $\xrightarrow{d}$ $p(h)$

$P(h=\text{true}) = \frac{1}{3}$
$P(h=\text{false}) = \frac{2}{3}$

- Enjoy Spring Break (next week)!
- Homework #3 submission closed.
- Project topic/papers due tonight 10 pm.
  - Submit the choice of topic and more than 5 selected papers in the Canvas discussion thread
  - 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/DATA/Segmentation_Data.zip**

Parametric Statistical Modeling

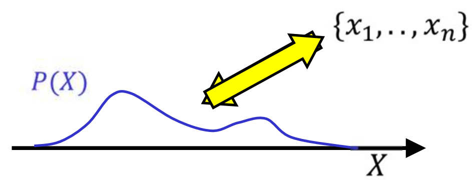
CSC 872

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?**



Non-Parametric Modeling: Review

- Histogram & Kernel Density Estimation KDE
  - *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 → model fitting / Regression
  - Basically fit the function to the data …in another words …
  - **Estimate the parameters** that make the function fit best to the data

New Strategy: Parametric Modeling

- Let's **use prior knowledge/assumption** of the target distribution !!!

- **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 α* i.i.d. change its shape & location
  - $\boldsymbol{x_1, x_2, .., x_N}$ ~ (i.i.d.) $\boldsymbol{p(x|\alpha) := f(x,\alpha)}$
  - **You know a form of *f* BUT you don't know the value of α**
- ***For what α are these samples most likely?***

$$\alpha^{mle} = argmax_{\alpha} \underbrace{p(\boldsymbol{x_1}, .., \boldsymbol{x_N}|\alpha)}_{\text{likelihood}}$$

MLE for Gaussian (Normal) Mean

- Suppose we have $\boldsymbol{x_1},..,\boldsymbol{x_N}$ **~** (i.i.d.) $N(\mu,\sigma^2)$
- But you don't know μ (known N and σ^2)
- MLE: For which μ is $\boldsymbol{x_1},..,\boldsymbol{x_N}$ most likely***?***

$$\mu^{mle} = \text{argmax}_\mu p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\mu,\sigma^2)$$

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

Algebra&Calculus to simplify the problem &

Solving $\nabla p(x) = 0$ to maximize the likelihood

Some Algebra

$X_1 \perp X_2 \longrightarrow P(X_1,X_2) = P(X_1)P(X_2)$

$$\mu^{mle} = \text{argmax}_\mu p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\mu,\sigma^2) \quad \text{MLE}$$

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

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

$$= \text{argmax}_\mu \sum_{n=1}^{N} \log[p(\boldsymbol{x_n}|\mu,\sigma^2)] \quad \log\prod = \sum\log$$

Log-likelihood !!!

$\log_e e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$= \text{argmax}_\mu \sum_{n=1}^{N} -\frac{(\boldsymbol{x_n}-\mu)^2}{\cancel{2\sigma^2}} + \cancel{C} \quad \text{Plugging in Gaussian}$$

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

$$= \text{argmin}_\mu \sum_{n=1}^{N} (\boldsymbol{x_n}-\mu)^2$$

Removing parts that are not related to the optimization

More Algebra

$$\boxed{\mu^{mle} = \text{argmax}_{\mu} p(\boldsymbol{x_1}, .., \boldsymbol{x_N} | \mu, \sigma^2)}$$

$$= \text{argmin}_{\mu} \sum_{n=1}^{N} (\boldsymbol{x_n} - \mu)^2$$

$$= \mu \quad \text{s.t. } 0 = \frac{\partial}{\partial \mu} \left[\sum_{n=1}^{N} (\boldsymbol{x_n} - \mu)^2 \right]$$

Argmin/Argmax is Solving ∇_{α} p(x,α) = 0

$$0 = -\sum_{n=1}^{N} \cancel{2}(\boldsymbol{x_n} - \mu^{mle})$$

$\sum_{n=1}^{N} \mu^{mle} = \sum_{n=1}^{N} x_n$

$\mu^{mle} \times N$

Do differentiation

$$\boxed{\mu^{mle} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x_n}}$$

Solve it about μ

What do we get?

- MLE μ^{mle} of a normal distribution is a **sample mean**

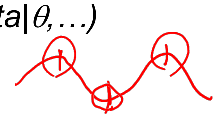
$$\mu^{mle} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- 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 θ assuming known form of $p(Data|\theta,\ldots)$

1. Derive **log-likelihood** (LL): LL = log $p(Data|\theta,\ldots)$
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 θ for maximizing $p(Data|\theta,\ldots)$
5. **Check if the solution is a maximum** instead of minimum or saddle point



CSC872: PAMI – Kazunori Okada (C) 2025 13

13

For more than one parameters

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

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 a set of equations by setting

$$\left\{\begin{array}{l} \partial LL/\partial\theta_1 = 0 \\ \partial LL/\partial\theta_2 = 0 \\ \quad\vdots \\ \partial LL/\partial\theta_n = 0 \end{array}\right.$$

4. **Solve the simultaneous equations** about $\theta=\theta_1,..,\theta_n$
5. **Check if the solution is a maximum**

CSC872: PAMI – Kazunori Okada (C) 2025 14

14

STOP: Back to MLE of Gaussian

$$\theta^{mle} = \text{argmax}_{\theta} p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\theta = (\mu, \sigma^2))$$

$$LL = \log[p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\mu, \sigma^2)]$$
$$= -0.5N\log 2\pi - 0.5N\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\boldsymbol{x_n} - \mu)^2$$

μ & σ^2?

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

$$\frac{\partial LL}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(\boldsymbol{x_n} - \mu)^2 = 0 \quad \rightarrow \quad \sigma^2_{mle} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x_n} - \mu^{mle})^2$$

Sample variance!!!

Unbiased Estimator

- Unbiased Estimator
  - **Expected value** of the estimate is the same as the **true value** of the estimate
- Suppose $\boldsymbol{x_1,..,x_N}$ ~ (i.i.d.) $\boldsymbol{N(\mu,\sigma^2)}$

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

$\sum_n \mu = \mu \sum_n 1 = \mu N$

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

$$\boxed{E[\mu^{mle}] = \mu}$$

μ^{mle} **is unbiased**

Biased Estimator

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

- Suppose $\boldsymbol{x_1,..,x_N}$ ~ (i.i.d.) $\boldsymbol{N(\mu,\sigma^2)}$

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

$$E[\sigma^2_{mle}] = E\left[\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x_n} - \mu^{mle})^2\right] = \frac{1}{N}E\left[\sum_n \boldsymbol{x_n^2} - \frac{1}{N}\sum_n\sum_{n'}\boldsymbol{x_n x'_n}\right]$$

$$= \frac{1}{N^2}E\left[(N-1)\sum_n \boldsymbol{x_n^2} - \sum_{n\neq n'}\boldsymbol{x_n x'_n}\right]$$

$$= \frac{N-1}{N^2}\sum_n E[\boldsymbol{x_n^2}] - NE[\boldsymbol{x_n}]^2 \boxed{= \frac{N-1}{N}\sigma^2}$$

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

$$\boxed{E[\sigma^2_{mle}] \neq \sigma^2}$$

⇒ σ^{mle} **is biased!!!**

CSC872: PAMI – Kazunori Okada (C) 2025 17

17

$$\sigma^2_{mle} = \frac{1}{N}\sum_n (x_n - \mu^{mle})^2 \qquad \mu^{mle} = \frac{1}{N}\sum_n x_n \qquad \sum_n (x_n - \mu)$$

$$= \sum x_n - \mu N$$

$$= N\left(\frac{\sum x_n}{N} - \mu\right)$$

$$= N(\mu^{mle} - \mu)$$

$$E[\sigma^2_{mle}] = \frac{1}{N}E\left[\sum_n (x_n - \mu^{mle})^2\right]$$

$$= \frac{1}{N}E\left[\sum_n (x_n - \mu + \mu - \mu^{mle})^2\right]$$

$$= \frac{1}{N}E\left[\sum_n \left((x_n-\mu)^2 + (\mu^{mle}-\mu)^2 - 2(x_n-\mu)(\mu^{mle}-\mu)\right)\right]$$

$$= \frac{1}{N}E\left[\sum_n (x_n-\mu)^2 + (\mu^{mle}-\mu)^2 N - 2(\mu^{mle}-\mu)\underline{\sum(x_n-\mu)}\right]$$

$$= \frac{1}{N}E\left[\sum_n (x_n-\mu)^2 + \underline{(\mu^{mle}-\mu)^2 N - 2(\mu^{mle}-\mu)^2 N}\right]$$

$$= \frac{1}{N}E\left[\sum_n (x_n-\mu)^2 - N\cdot(\mu^{mle}-\mu)^2\right] \qquad \boxed{E(\mu^{mle}-\mu)^2 = \frac{\sigma^2}{N}}$$

$$= \frac{1}{N}\left[\sum_n E(x_n-\mu)^2 - N\cdot E(\mu^{mle}-\mu)^2\right]$$

$$= \frac{1}{N}\left[\sigma^2\sum_n - N\cdot\frac{\sigma^2}{N}\right] = \frac{1}{N}(N\sigma^2 - \sigma^2) = \frac{N-1}{N}\sigma^2$$

CSC872: PAMI – Kazunori Okada (C) 2025 http://www.ee.columbia.edu/~dliang/files/mle_biased.pdf 18

18

What we gonna do?

- Bias: E[θ^{mle}] - θ

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

- Unbiased estimator from a biased one

$$E[\sigma^2_{mle}] = \frac{N-1}{N}\sigma^2 \qquad E[\sigma^2_{unbiased}] = \sigma^2 \qquad \frac{N}{N-1}\frac{1}{N}\sum_n (x_n-\mu)^2$$

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

- Is unbiased estimator always better?

$$\mu^{est1} = \boldsymbol{x_4}$$

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

- Asymptotically unbiased estimator

$$E[\sigma^2_{mle}] = \frac{N-1}{N}\sigma^2 \xrightarrow[N\to\infty]{} \sigma^2$$

And more … what if

- <u>TASK: Find θ assuming known form of $p(Data|\theta=\theta_1,..,\theta_n)$</u>

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

$$\begin{cases} \partial LL/\partial\theta_1 = 0 \\ \partial LL/\partial\theta_2 = 0 \\ \vdots \\ \partial LL/\partial\theta_n = 0 \end{cases}$$

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 $\partial LL/\partial\theta_1=0,..,\partial LL/\partial\theta_n=0$
- Oh no …
- But there is a savior …
- **Go Iterative !!!**

  Optimization

  (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

21

Maximum A Posteriori Estimation

Remember our Bayesian inference lecture. Treating a likelihood as a posterior yielded a wrong inference. In the same sense, we should be using a posterior for our parameter estimation problem!

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

$$\theta^{mle} = \text{argmax}_\theta p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\theta)$$

$$\theta^{map} = \text{argmax}_\theta p(\theta|\boldsymbol{x_1},..,\boldsymbol{x_N})$$

$$= \text{argmax}_\theta \frac{p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\theta)p(\theta)}{\int_{\theta'} p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\theta')p(\theta')\,d\theta'}$$

$$= \text{argmax}_\theta p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\theta)\underline{p(\theta)}$$

You need to provide a prior distribution

22

MAP for Gaussian (Normal) Mean

$e^a \times e^b = e^{a+b}$

- Suppose we have $\boldsymbol{x_1,..,x_N}$ ~ (i.i.d.) $\boldsymbol{N(\mu,\sigma^2)}$
- But you don't know μ
- **MAP**: Which μ maximizes posterior $\boldsymbol{p(\mu \,|x_1,..,x_N,\sigma^2)}$
- Set **the prior** also as a Gaussian $\boldsymbol{N(\mu_0,\sigma_0^2)}$

$\sigma_0^2 = 0$ $\quad \mu_0$ $\quad \sigma_0^2$ $\quad \mu_0$ $\quad \sigma_0^2 = \infty$ $\quad \mu_0$

$$\mu^{map} = \text{argmax}_\mu p(\mu|\boldsymbol{x_1},..,\boldsymbol{x_N},\sigma^2)$$

Bayes

$$= \text{argmax}_\mu p(\boldsymbol{x_1},..,\boldsymbol{x_N}|\mu,\sigma^2)p(\mu)$$

iid

$$= \text{argmax}_\mu N(\mu;\mu_0,\sigma_0^2)\prod_{n=1}^{N} N(\boldsymbol{x_n};\mu,\sigma^2)$$

$$= \text{argmax}_\mu N(\mu;\mu_1,\sigma_1^2) \qquad \mu_1 = \frac{\sigma^2\mu_0+\sigma_0^2\sum_{n=1}^N \boldsymbol{x_n}}{\sigma^2+\sigma_0^2 N}$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2 N}$$

MLE recipe: log+∇p=0

$$= \frac{\sigma^2\mu_0+\sigma_0^2\sum_{n=1}^N \boldsymbol{x_n}}{\sigma^2+\sigma_0^2 N}$$

$\sigma_0^2 \to 0$ very sure $\mu^{map} \approx \mu_0$

$\sigma_0^2 \to \infty$ very unsure $\mu^{map} \approx \mu^{mle}$

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 nuisance (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

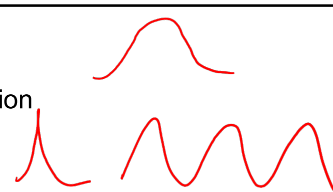
PF: Is it learning?

Bayesian 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