PARAMETRIC ESTIMATION

Shingchern D. You

Parametric estimation

- Assuming data samples are from a probability model $p(\boldsymbol{x}|\theta)$, want to estimate θ from samples
  - Example: Tossing a coin, $\theta = \mathrm{P}(\{\mathrm{H}\})$
  - Example: Male heights modeled in Gaussian, $\theta = \{\mu, \sigma\}$
- Knowing parameters help to make predictions
  - Example: Predict which face shown on next coin-tossing with θ

Parametric estimating methods

- Maximum likelihood estimation (MLE)
- Maximum a posteriori (MAP) estimation
- Bayes estimator (not covered)

MLE concept

- Likelihood (function) of θ given a set of samples $\mathcal{X}=\{x_i\}$ (samples from iid RV)

$$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_i p(x_i|\theta)$$

- Want to find $\hat{\theta}$ such that $l(\theta|\mathcal{X})$ is max, i.e.,

$$\hat{\theta} = \operatorname{argmax} l(\theta|\mathcal{X})$$

- Sometimes it is easier to find $\hat{\theta}$ by log likelihood

$$\mathcal{L}(\theta|\mathcal{X}) = \log p(\mathcal{X}|\theta) = \sum_i \log p(x_i|\theta)$$

MLE example

- Instead of finding closed form, we use examples
- Tossing a coin 5 times with results H, H, T, T, H
- What $\theta = P(\{H\})$ value can maximize the likelihood

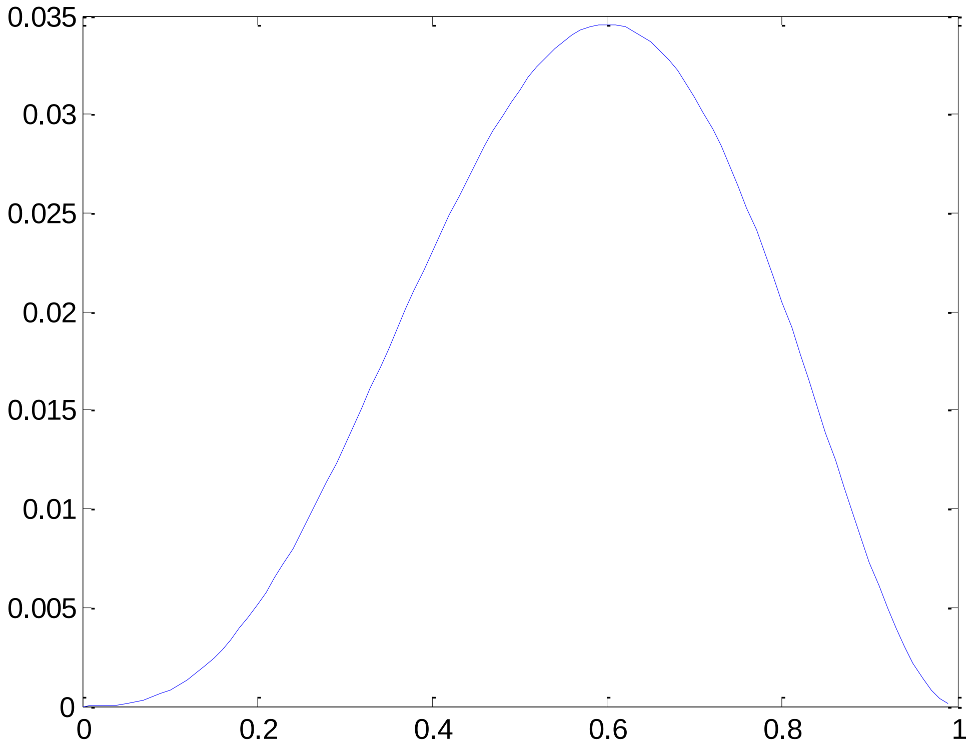
https://gurmeet.net/Images/puzzles/coin_toss_guess.png

MLE example

- Use this equation: $\prod_i p(x_i|\theta)$
- If $P(\{H\}) = \theta = 0.1$, we have $P(\{H,H,T,T,H\}) =$ $0.1*0.1*0.9*0.9*0.1 = 8.1 \times 10^{-4}$
- If $P(\{H\}) = \theta = 0.2$, we have $P(\{H,H,T,T,H\}) =$ $0.2*0.2*0.8*0.8*0.2 = 5.1 \times 10^{-3}$
- Repeat many times with different values
- Easier with a program

MLE example

- With a program for different θ, we have a plot
- $\theta = 0.6$ yields highest probability, i.e., $\hat{\theta} = 0.6$



MLE example

- Theoretic answer: $\hat{\theta} = \frac{N_H}{N_T}$
  - N_T is total tossing
  - N_H is tossing with head shown
- Derived by math (omitted)
- This result is widely used

Sample mean & sample variance

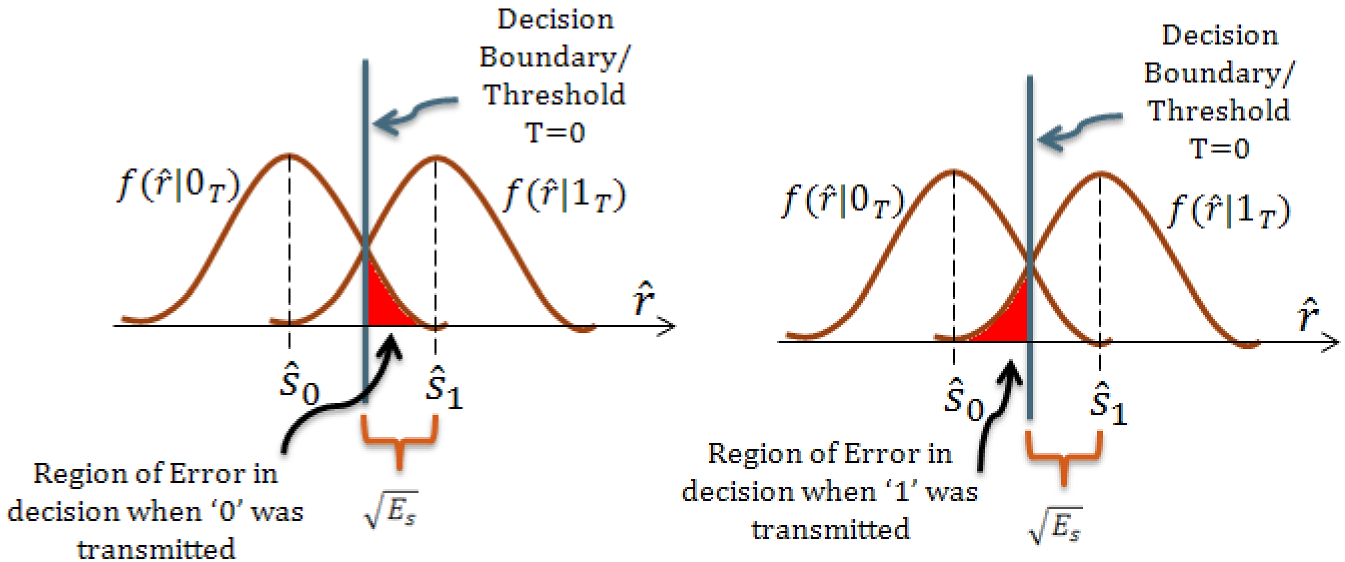
- Let x_i be samples from iid Gaussian
- Sample mean & sample variance are MLE of true mean and true variance

$$m = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - m)^2}{N}$$

Classification with MLE

- Noisy BPSK signal (**equal variance**, why)
- Let estimated means for 0 and 1 be m_0 & m_1
- Unknown symbol x is 0 if $|m_0 - x| < |m_1 - x|$



https://www.gaussianwaves.com/gaussianwaves/wp-content/uploads/2012/07/Calculating-Error-Probability.png

Classification with MLE

- We can extend the situation to
  - Unequal variance
  - Unequal probability of symbols 0 and 1
- Exercise

Bias & variance of estimators

- Let $\mathcal{X}=\{x_i\}$ be a set of samples with unknown parameter θ
- To do analysis, treat x_i as iid RV
- Let $d = d(\mathcal{X})$ be an estimator of θ
  - Example: equation to compute sample mean is an estimator
- Bias of estimator $= E[d(\mathcal{X})] - \theta$, where $E[\cdot]$ denotes expectation
- Variance of estimator $=E[(d - E[d])^2]$

Bias & variance of estimators

- Mean square error (MSE) $= E[(d-\theta)^2] = E[(d-E[d])^2] + (E[d]-\theta)^2$
- 1st term is variance
- 2nd term is square of bias
- To reduce MSE, we need to reduce both

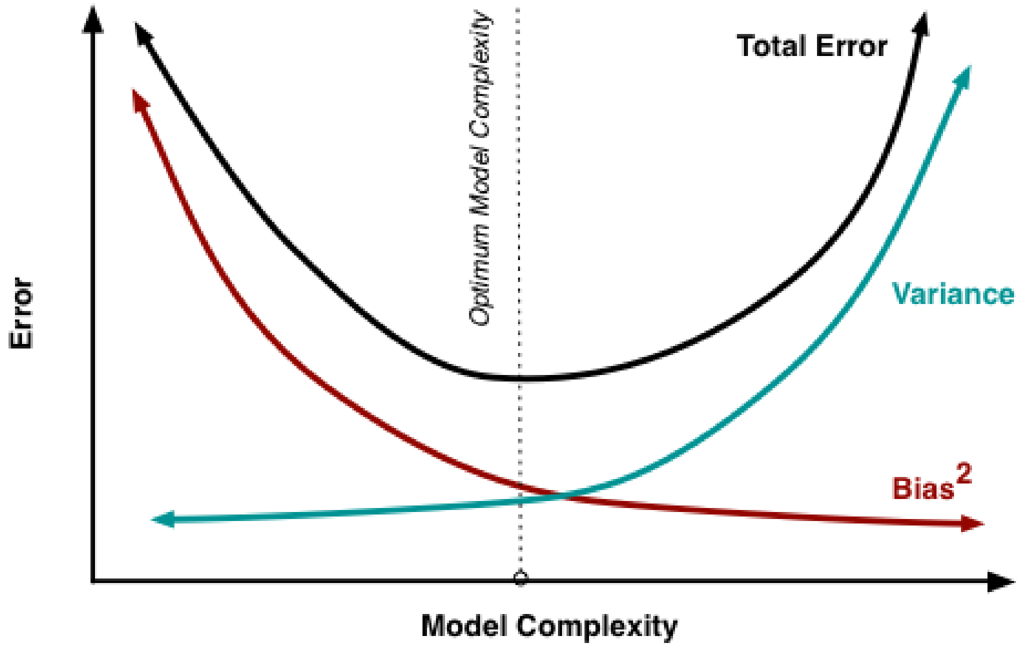
Bias & variance of classifiers

- We can similarly define bias & variance of a classifier
- Detailed math is omitted here
- A good ref is here: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias & variance of classifiers

- Conceptual results of increasing model complexity

(from http://scott.fortmann-roe.com/docs/BiasVariance.html)



What does model complexity mean

- In k-NN case, k controls complexity
- In neural networks, number of trainable (connection) weights
- In SVM, order of polynomial kernel
- In BPSK case
  - Estimate mean values
  - Estimate probability of symbol 0 and symbol 1

Rethinking ML estimation

- This equation seems counter-intuitive…
- Throwing divination blocks (擲筊杯) 15 times with 15 “agrees,” what is probability of “agree” shown on next throwing

https://img.ruten.com.tw/s1/a/9b/65/21917183440741_479.jpg

MAP (Maximum a Posteriori)

- ML says 1.0, but we know it is likely 0.5 because tossing divination blocks is modeled as "repeated" independent trials
- That is the difference between ML and MAP
- ML is derived ONLY based on observation
- MAP incorporates ***prior*** knowledge into estimation
- Recall Bayes theorem

$$P(\theta|\chi) = \frac{P(\chi|\theta)P(\theta)}{P(\chi)}$$

MAP

- MAP estimator wants to find

$$\theta_{MAP} = \arg\max_{\theta} \mathrm{P}(\theta|\boldsymbol{\chi})$$

- As $P(\boldsymbol{\chi})$ does not affect the max operation, we need to consider only

  $P(\boldsymbol{\chi}|\theta)P(\theta)$

- The first term is equal to ML, the second term is the ***a priori* probability**

MAP

- The probability $P(\{\mathrm{H}\}) = \theta$ is typically modeled as outcome from beta distribution, whose pdf is

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{(a-1)}(1-x)^{(b-1)}$$

- Therefore, we need to determine values of a and b (*a priori* knowledge) in order to use MAP

Numerical solution to MAP

- Generate many uniformly spaced values for θ, say, 100 numbers between 0 and 1
- Compute likelihood for each θ, OK to use pdf in place of $P(\theta)$
- Find θ_{max}
- With lots of math, we have

$$\theta_{MAP} = \frac{N_H + a - 1}{N + a + b - 2}$$

Source: www.mi.fu-berlin.de/wiki/pub/ABI/Genomics12/MLvsMAP.pdf

MAP example

- To have a “mean” $\theta = 0.5$, we set a = b
- Use our coin-tossing example
- If a = b = 1, we have $\theta_{MAP} = 0.6$ (same as ML)
- If we are more confident about the *prior* knowledge, we can set larger values of a and b, such as a = b = 10

MAP example

- If a = b = 10, we have $\theta_{MAP} = 0.522$