(Univariate) Parametric Methods

Summary:

1. Choose a model
2. Use observed data to estimate model parameters
3. Use the trained model to make predictions

Outline

- How do we estimate model parameters from data?
  - Maximum Likelihood Estimation
  - Bayesian estimation
- Bias and Variance
  - Estimators
  - Models

Parametric Methods

We want to know the distribution of our data, $p(x)$

- Assumption: The distribution has a particular form (e.g. Gaussian)
- We want to estimate its parameters (e.g. μ, σ^2) from the data
- We'll look at two approaches:
  - Maximum Likelihood Estimation
  - Bayesian estimation

In other words..

- We have a sample $X = \{x^t\}_{t=1}^{N}$ where $x^t \sim p(x)$
- Assume a form for $p(x|\theta)$
  - θ : sufficient statistics
- Estimate θ using X
- E.g., $x^t \sim N(\mu, \sigma^2)$ where $\theta = \{\mu, \sigma^2\}$

Maximum Likelihood Estimation

Classical approach: Choose the parameter value that maximizes the probability of the observed data

- Probability of a particular data point x^t:
  $p(x^t|\theta)$

Maximum Likelihood Estimation

Classical approach: Choose the parameter value that maximizes the probability of the observed data

- Probability of a particular data point x^t:
  $p(x^t|\theta)$
- Likelihood of θ given the sample X
  $l(\theta|X) = p(X|\theta) = \prod_t p(x^t|\theta)$

This is what we want to maximize

- Working with logs instead:
- Log likelihood
  $L(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)$

Maximum Likelihood Estimation

- Working with logs instead:
- Log likelihood
  $L(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)$

- Find the maximum:
- Maximum likelihood estimator (MLE)
  $\theta^* = \text{argmax}_\theta \, L(\theta|X)$

Example: Bernoulli

- Bernoulli: Two states, failure/success, $x \in \{0, 1\}$

$$P(x) = p^x(1-p)^{(1-x)}$$
$$L(p|X) = log \prod_t p^{x^t}(1-p)^{(1-x^t)}$$
$$\text{MLE: } p = \sum_t x^t/N$$

- Bernoulli: Two states, failure/success, $x \in \{0, 1\}$

$$
\begin{aligned}
P(x) &= p^x(1-p)^{(1-x)} \\
L(p|X) &= log \prod_t p^{x^t}(1-p)^{(1-x^t)} \\
\text{MLE: } p &= \sum_t x^t/N
\end{aligned}
$$

- For multinomial:

$$
\text{MLE: } p_i = \sum_t x_i^t/N
$$

- Normally distributed sample: $X \sim N(\mu, \sigma^2)$
- Density function:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

- MLE for μ and σ^2 :

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t \left(x^t - m\right)^2}{N}$$

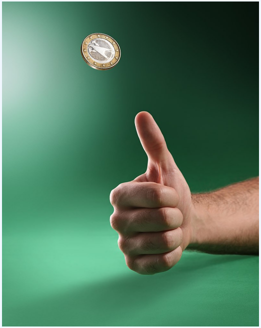
Shortcomings of MLE

- Usually no analytical solutions to more complicated distributions
- Sensitive to too little data and non-repeated events
- sometimes the most likely is not the most probable

- Coin Toss example:

$$\boldsymbol{X} = \{1, 1, 1\}$$
$$\hat{p} = \frac{3}{3} = 1$$

- How can we improve this?



Frequentist vs Bayesian statistics

Frequentist vs Bayesian statistics

Frequentist

- Probabilities as frequencies of occurrences
  (How often does the coin land Heads up?)
- Parameters have one true value which we estimate
- Estimates are based on the data (sample) only

Bayesian

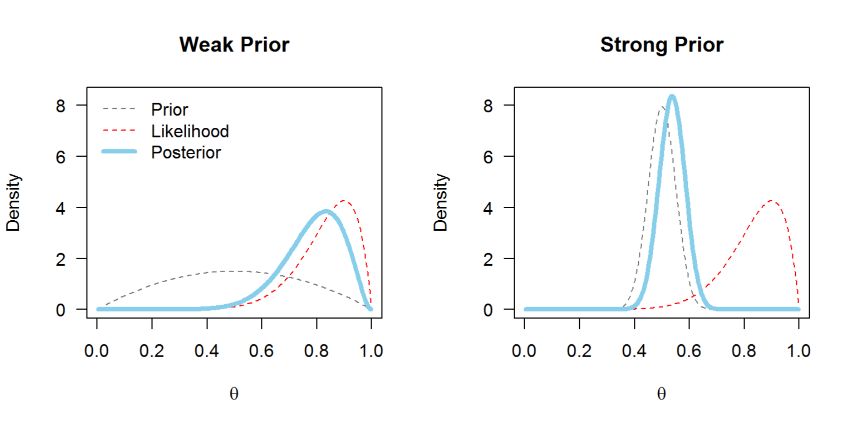
- Probabilities as degree of certainty
  (How certain are we that Brexit will happen?)
- Parameters are themselves *random variables* with probability distributions
- Estimates incorporate *prior knowledge*

- θ is a random variable with prior $p(\theta)$ that describes our beliefs before seeing the data
- Bayes' rule gives the posterior distribution:

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})} = \frac{p(\mathcal{X}|\theta)p(\theta)}{\int p(\mathcal{X}|\theta')p(\theta')d\theta'}$$

How likely parameter values are after seeing the data

Coin example: we toss 10 times and get 9/10 Heads



jimgrange.wordpress.com/2016/01/18/pesky-priors/

Bayes estimator

Resulting estimate for the probability at a new sample point

$$p(x|X) = \int p(x|\theta)p(\theta|X)d\theta$$

Rather than using the prediction of a single parameter value ('frequentist statistics'), we average the prediction of every parameter value using its posterior distribution ('Bayesian statistics').

Bayes estimator

To simplify, the posterior distribution is often reduced to a point estimate

- Maximum a Posteriori (MAP):

  $$\theta_{MAP} = \mathrm{argmax}_\theta\, p(\theta|X)$$

  (If the prior is flat, this will be the same as the Maximum Likelihood (ML) estimate:

  $$\theta_{ML} = \mathrm{argmax}_\theta\, p(X|\theta)$$

- Bayes' estimator, :

  $$\theta_{Bayes'} = E[\theta|X] = \int \theta p(\theta|X) d\theta$$

These work best when posterior distributions are unimodal with a narrow peak

- Air France AF 447, Brazil - France, June 2009
- Disappeared over the ocean in a storm

P Kierkowski

Bayesian statistics: The search for AF 447

- Intensive search efforts
  - Aerial search
  - Multiple search vessels
  - Nuclear submarine
  - Sonar arrays
  - Robotic submarines
- Scanned > 1 million km^2 of ocean
- Found > 600 items of debris
- Search went on for a year ..but wreckage not found

Bayesian statistics: The search for AF 447

- Intensive search efforts
  - Aerial search
  - Multiple search vessels
  - Nuclear submarine
  - Sonar arrays
  - Robotic submarines
- Scanned > 1 million km^2 of ocean
- Found > 600 items of debris
- Search went on for a year ..but wreckage not found

Statistical Science
2014, Vol. 29, No. 1, 69–80
DOI: 10.1214/13-STS420

Search for the Wreckage of Air France Flight AF 447[1]

Lawrence D. Stone, Colleen M. Keller, Thomas M. Kratzke and Johan P. Strumpfer

Bayesian statistics: The search for AF 447

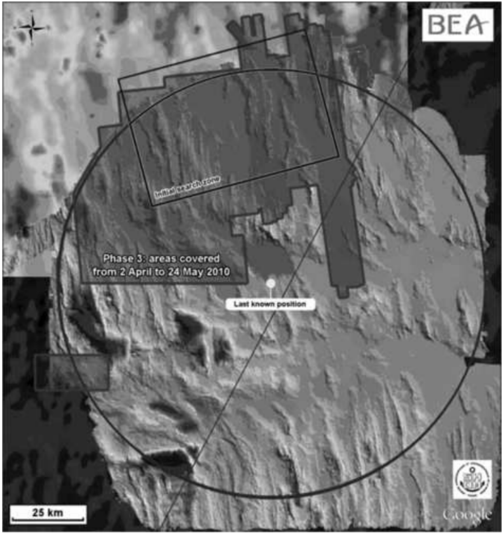


Fig. 6. *Regions searched by active side-looking sonar in April–May 2010.*

Bayesian statistics: The search for AF 447

SEARCH FOR AIR FRANCE FLIGHT AF 447

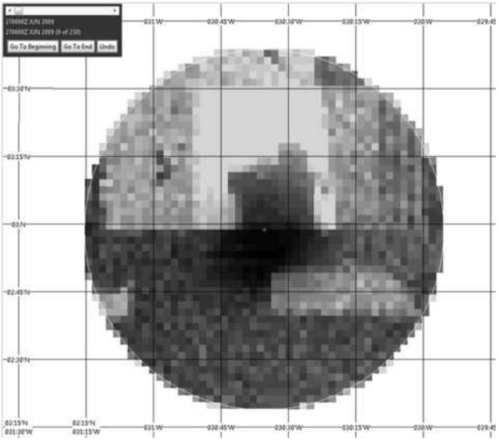
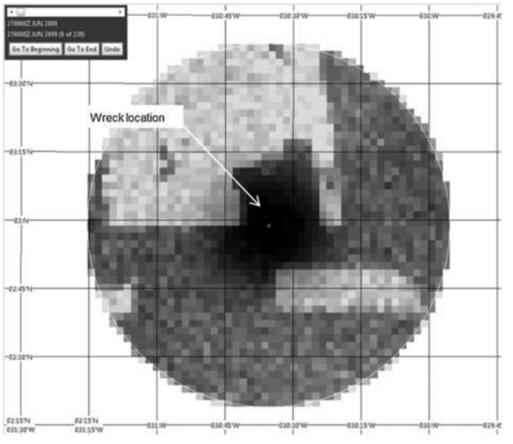


Fig. 7. *Posterior distrubtion* $\tilde{P}_{1234}$.

Found after one week!

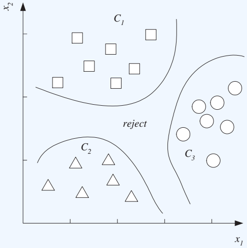


Bayesian Parameter Estimates

- Pros: works well when the sample size N is small (if the prior is helpful).
- Cons: computationally harder (needs to compute, usually approximately, integrals or summations). Needs to define a prior.

So we found our MLE of θ, how do we use that?

- From last lecture: Discriminant functions for classification:
  $g_i(x) = p(x|C_i)P(C_i)$
  or
  $g_i(x) = \log p(x|C_i) + \log P(C_i)$



- $g_i(x) = \log p(x|C_i) + \log P(C_i)$
- Assuming $p(x|C_i) \sim N(\mu, \sigma^2)$:

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Parametric Classification

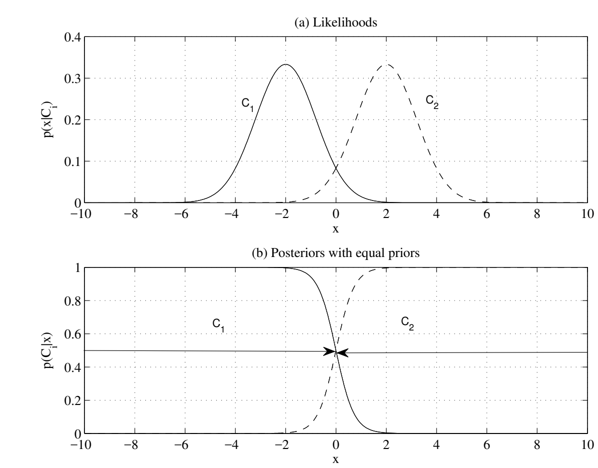
- Given the sample: $X = \{x^t, r^t\}_{t=1}^N$
  $X \in R$
  $r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j,\ j \neq i \end{cases}$
- ML estimates are
  $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$
  $m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$
  $s_i^2 = \frac{\sum_t (x_t - m_i)^2 r_i^t}{\sum_t r_i^t}$
- Discriminant
  $g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$

Parametric Classification

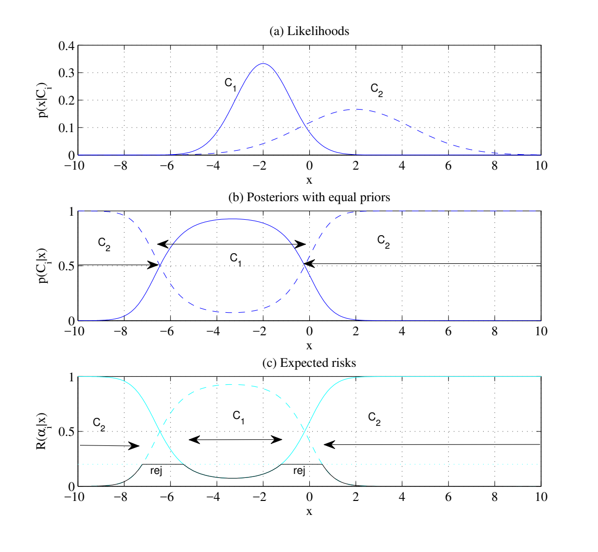
- Discriminant
  $g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$
- Constant term can be dropped
  $g_i(x) = -\log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$
- If equal priors: $g_i(x) = -\log s_i - \frac{(x-m_i)^2}{2s_i^2}$
- If equal variance: $g_i(x) = -(x-m_i)^2$

Parametric Classification

With equal prior and variance, assign to nearest mean.
The decision boundary is the midpoint between the means.

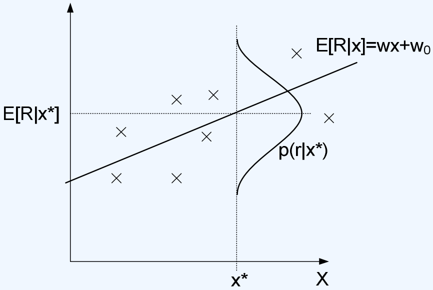


With unequal variances, two thresholds:



Regression

- $r = f(x) + \epsilon$
- $\epsilon \sim N(0, \sigma^2)$
- Estimator: $g(x|\theta)$
- $p(r|x) \sim N(g(x|\theta), \sigma^2)$



- Log likelihood:

$$\begin{aligned} L(\theta|X) &= \log \prod_{t=1}^{N} p(x^t, r^t) \\ &= \log \prod_{t=1}^{N} p(r^t|x^t) + \log \prod_{t=1}^{N} p(x^t) \end{aligned}$$

last term can be ignored

- Insert $p(r|x) \sim N(g(x|\theta), \sigma^2)$ in the log-likelihood function

$$L(\theta|X) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-[r^t - g(x^t|\theta)]^2/2\sigma^2}$$
$$= -N \log \sqrt{2\pi\sigma} - \frac{1}{\sqrt{2\sigma^2}} \sum_{t=1}^{N} [r^t - g(x^t|\theta)]^2$$

- Ignore terms independent of θ and maximize:

$$-\frac{1}{2} \sum_{t=1}^{N} [r^t - g(x^t|\theta)]^2$$

Regression: From LogL to Error

- Maximizing this:

$$-\frac{1}{2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

- is the same as minimizing the least squares estimate:

$$E(\theta|X) = \frac{1}{2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

How do we minimize?

$$E(\theta|X) = \frac{1}{2}\sum_{t=1}^{N}[r^t - g(x^t|\theta)]^2$$

- In case of linear regression, our model is: $g(x^t|w_1, w_0) = w_1 x^t + w_0$
- We want to find the least squares estimates of w_1, w_0
- Take the derivative wrt. w_1 and w_0 and set to 0

- Two equations in two unknowns

  $\sum_t r^t = Nw_0 + w_1 \sum_t x^t$

  $\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t (x^t)^2$
- In vector matrix form:

  $\mathbf{Aw} = \mathbf{y}$

Linear Regression

- In vector matrix form:

  $\mathbf{A}\mathbf{w} = \mathbf{y}$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

- Solved as:

  $\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$

- Polynomial regression has the form:
  $g(x^t|w_k, \ldots, w_2, w_1, w_0) = w_k(x^t)^k + \ldots + w_2(x^t)^2 + w_1x^t + w_0$
- The model is still linear in the parameters
- The least squares estimate still has the form
  $\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$

Bias and Variance

Bias and Variance

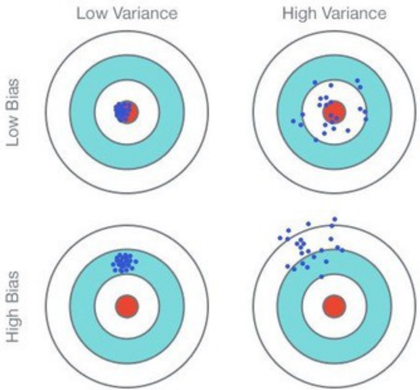


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

Bias and Variance of an estimator

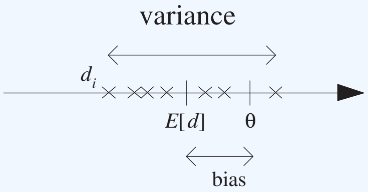
- We have a sample X from some population
- From X, we calculate some statistic $d(X)$ (could be the mean, maximum, etc)
- $d(X)$ is itself a random variable (because the sample X is a random set)

- We have a sample X from $p(x;\theta)$
- We have an estimator of θ: $d = d(X)$
- How good is our estimator?
  Mean square error:

$$r(d,\theta) = E[(d(X) - \theta)^2]$$

Bias and Variance of an estimator

- Bias of the estimator: $b_\theta(d) = E[d(X)] - \theta$
- Variance of the estimator: $E[(d(X) - E[d(X)])^2]$



The overall error of the estimator can be decomposed:

Mean square error:

$$\begin{aligned} r(d,\theta) &= E[(d-\theta)^2] \\ &= (E[d]-\theta)^2 + E[(d-E[d])^2] \\ &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

- We are not just interested in parameter estimates
- We want to know:
  How good can we expect our trained model to be!

- In a regression setting: $r = f(x) + \epsilon$
- Our regression estimate: $g(x)$

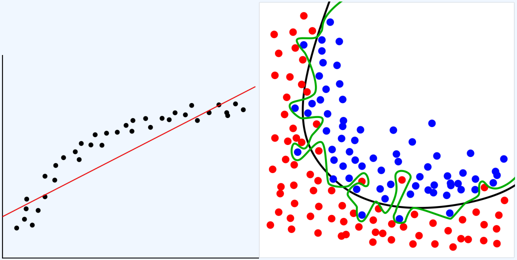
$$E[(r - g(x))^2|x] = \underbrace{E[(r - E[r|x])^2|x]}_{noise} + \underbrace{(E[r|x] - g(x))^2}_{squared\ error}$$

$$E_X[(E[r|x] - g(x))^2|x] = \underbrace{(E[r|x] - E_X[g(x)])^2}_{bias} + \underbrace{E_X[(g(x) - E_X[g(x)])^2]}_{variance}$$

Full decomposition:
https://www.youtube.com/watch?v=zUJbROoWavo

Bias/Variance Dilemma

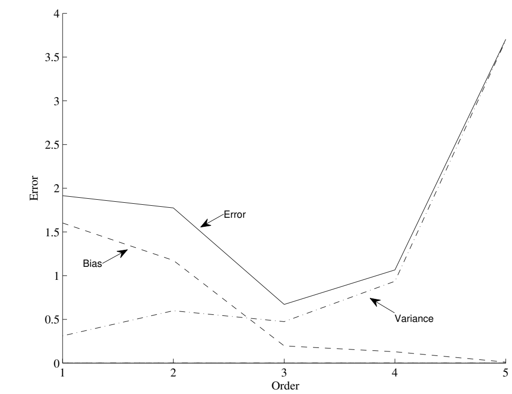
- Simple models (underfitting):
  - Low variance (don't change a lot with new sample)
  - High bias (can't capture the true relations)
- Complex models (overfitting):
  - High variance (memorizes training data)
  - Low bias (can find complex relations)



towardsdatascience.com

Bias/Variance Dilemma

- As we increase complexity:
  - bias decreases (a better fit to data)
  - variance increases (fit varies more with data)



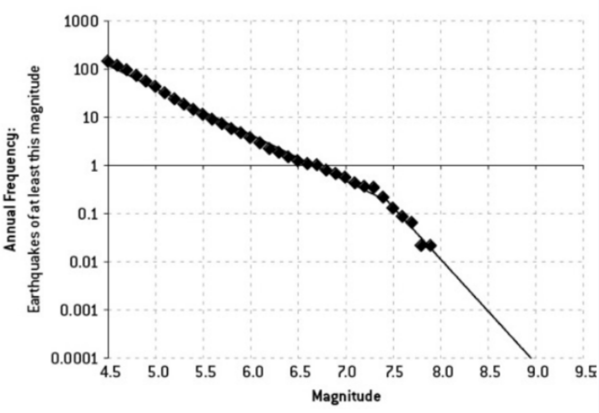
Overfitting: The Fukushima Disaster

- Fukushima nuclear power plant
- Destroyed by earthquake-triggered tsunami on 2011

Digital Globe

Overfitting: The Fukushima Disaster

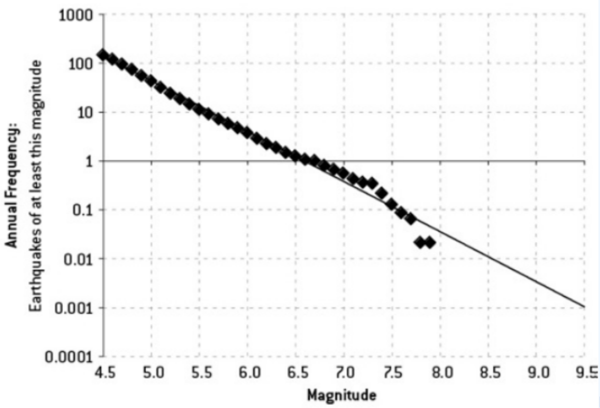
- Designed to withstand 8.6 Richter earthquake, 5.7m tsunami
- Safety analysis assumed non-linear fit



B. Stacey 2015

Overfitting: The Fukushima Disaster

- Gutenberg-Richter law: logarithmic
- 2011 Earthquake: 9.0 Richter, 14m tsunami
-



B. Stacey 2015