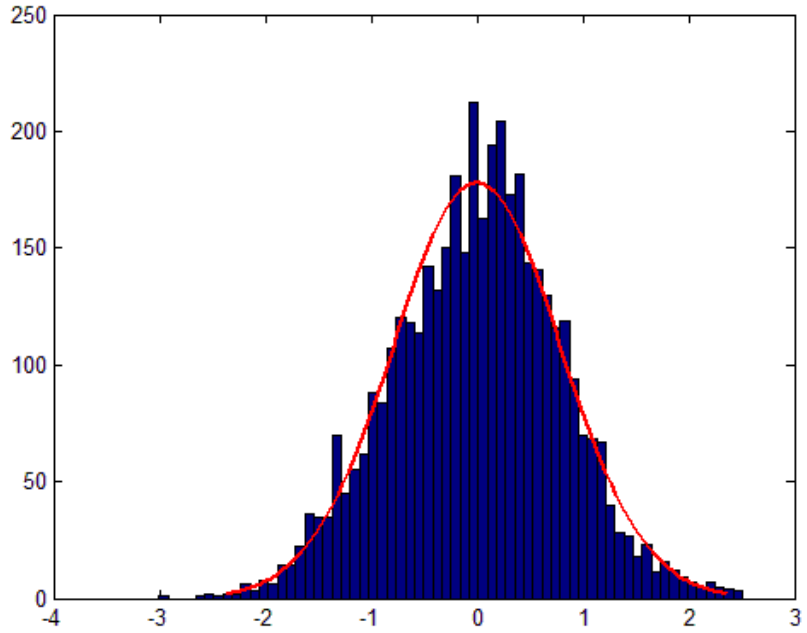
Simple Parameter and Density Estimation



Ariel Shamir

Zohar Yakhini

IDC HERZLIYA

Types of Learning Tasks

- Regression
  - Given $\{x_i, y_i\}$ find f such that $y = f(x)$
- Classification
  - Given $\{x_i, y_i\}$ where $y_i \in \{0,1\}$ for training, determine for a new x if $x \in C_0$ or $x \in C_1$
- **Density Estimation**
  - **Given $\{x_i\}$ find PDF that best explains the data**
- Clustering
  - Given $\{x_i\}$ find a partition to k subsets under some constraints

MAP classification (next week)

Classify an instance with observed properties $\vec{x}$ as

$$\underset{i}{\operatorname{argmax}}\, P(\vec{x}|A_i)P(A_i)$$

Outline of next steps

$$\text{argmax}_i\{P(x|A_i)P(A_i)\}$$

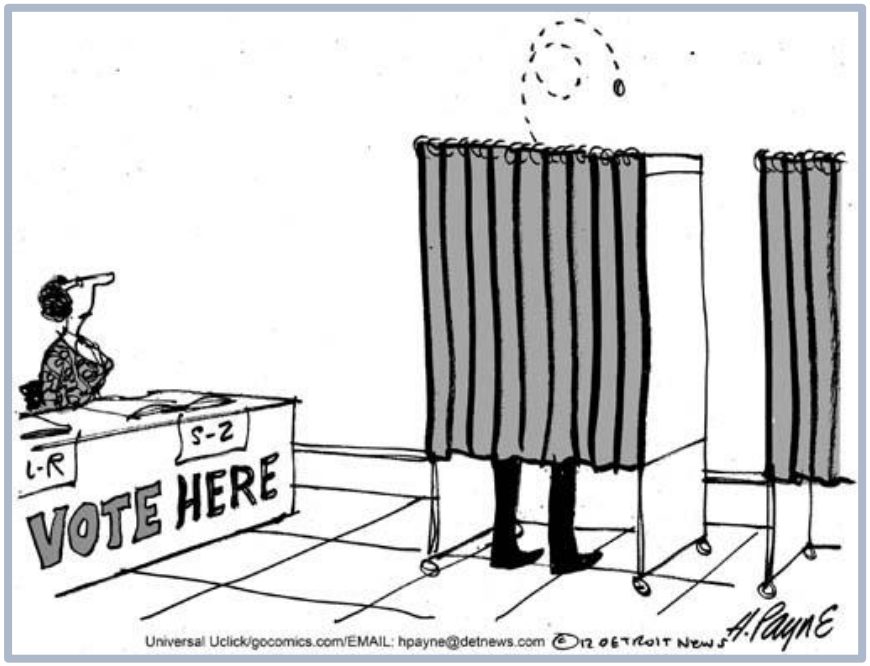
Need to estimate $P(x|A_i)$ and $P(A_i)$

- Estimating Probabilities and Densities: parametric vs. non-parametric
- Estimating in 1D vs. High-Dim
- Statistical dependence and conditional independence
- Naïve Bayes classifiers
- EM algorithm

MLE

- A straightforward approach to parameter estimation.
- Directly works in simple cases.
- Forms the basis for most parameter estimation approaches

An easy case: Coin tossing



- Assuming
  - + A coin has a probability *p* of being heads (H), *q=1-p* of being tails (T).
  - + Observation: We toss the coin m times, observing a set of Hs and Ts.

- What is the (most likely) value of *p*?
  → <u>Use the MLE principle, given the observation</u>

Coin tossing … cont

What would have happened if we used a binomial model??

Assume that we tossed N times and observed m results of H

$$L(\Theta) = \log P(D \mid \Theta) = \log p^m (1-p)^{N-m}$$
$$= m \log p + (N-m)\log(1-p)$$

$$\frac{dL(\Theta)}{dp} = \frac{d(m \log\ p + (N-m)\log(1-p))}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$

$$\Rightarrow \quad \hat{p} = m/N$$

Confidence interval for proportions

$$Prob\left(p \in \left[\hat{p} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \approx 1-\alpha$$

Steps of MLE

- Given
  - A set of observed values
    $$D = \{x_1, \ldots, x_m\}$$
  - A model and candidate vector of parameters for this model, θ
- We define
  - The likelihood of the candidate model given the data,
    $$\Lambda(\theta|D) = P(D \mid \theta) .$$
  - The Log-likelihood of the model given the data:
    $$L(\theta) = \log P(D|\theta)$$
- In MLE we seek

$$\theta^* = \underset{\theta \in \Omega}{\operatorname{argmax}} L(\theta)$$

MLE for independent identically distributed instances

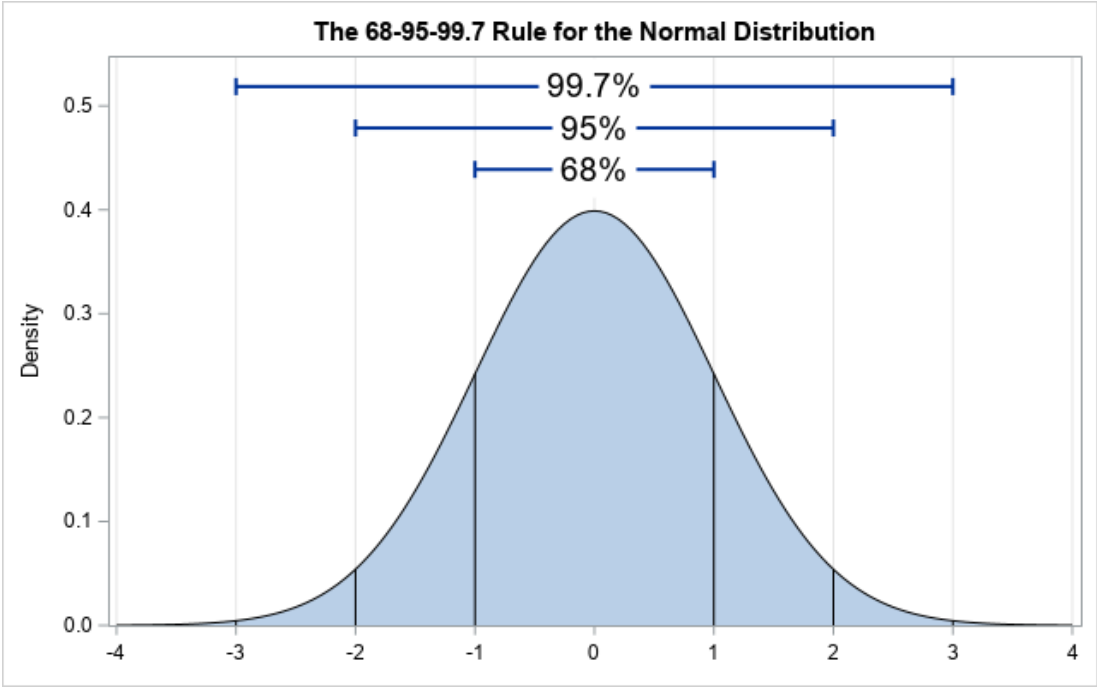
- We often assume that the data instances/observations are a result of independent identically distributed (i.i.d.) random variables.
  This is the same as assuming independent repeats of the same generation mechanism.
- The above then becomes:

$$\theta^* = \underset{\theta \in \Omega}{\operatorname{argmax}} \sum_{i=1}^{m} \log P(x_i|\theta)$$

- Solving the optimization problem varies in its complexity, depending on the form of $P(x|\theta)$ – the (common) pdf of the underlying variables

Normal Distribution

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



- Normal distributions are determined by two parameters: μ and σ.
- Given m observed values of a variable X, we want to estimate the mean and variance of its normal approximation:

A useful fact

If $X \sim N(\mu, \sigma)$

then

$Z = (X - \mu)/\sigma$

is a standard normal random variable

You can think of Z as measuring, for every instance of X drawn, the distance of the obtained value, from the expected value, in units of standard deviations

Steps of MLE

- Given
  - A set of observed values
    $$D = \{x_1, \ldots, x_m\}$$
  - A model and candidate vector of parameters for this model, θ
- We define
  - The likelihood of the candidate model given the data,
    $$\Lambda(\theta|D) = P(D \mid \theta) .$$
  - The Log-likelihood of the model given the data:
    $$L(\theta) = \log P(D|\theta)$$
- In MLE we seek

$$\theta^* = \underset{\theta \in \Omega}{\operatorname{argmax}} L(\theta)$$

Normal MLE, cont

We observe measurements $D = \{x_1, \ldots, x_n\}$ which we assume to be independent and to be coming from some normal distribution.

In this case $\theta = (\mu, \sigma^2)$

And

$$\Lambda(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Its easier to work with the log-likelihood:

$$L(\theta) = -n \ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Normal MLE, cont

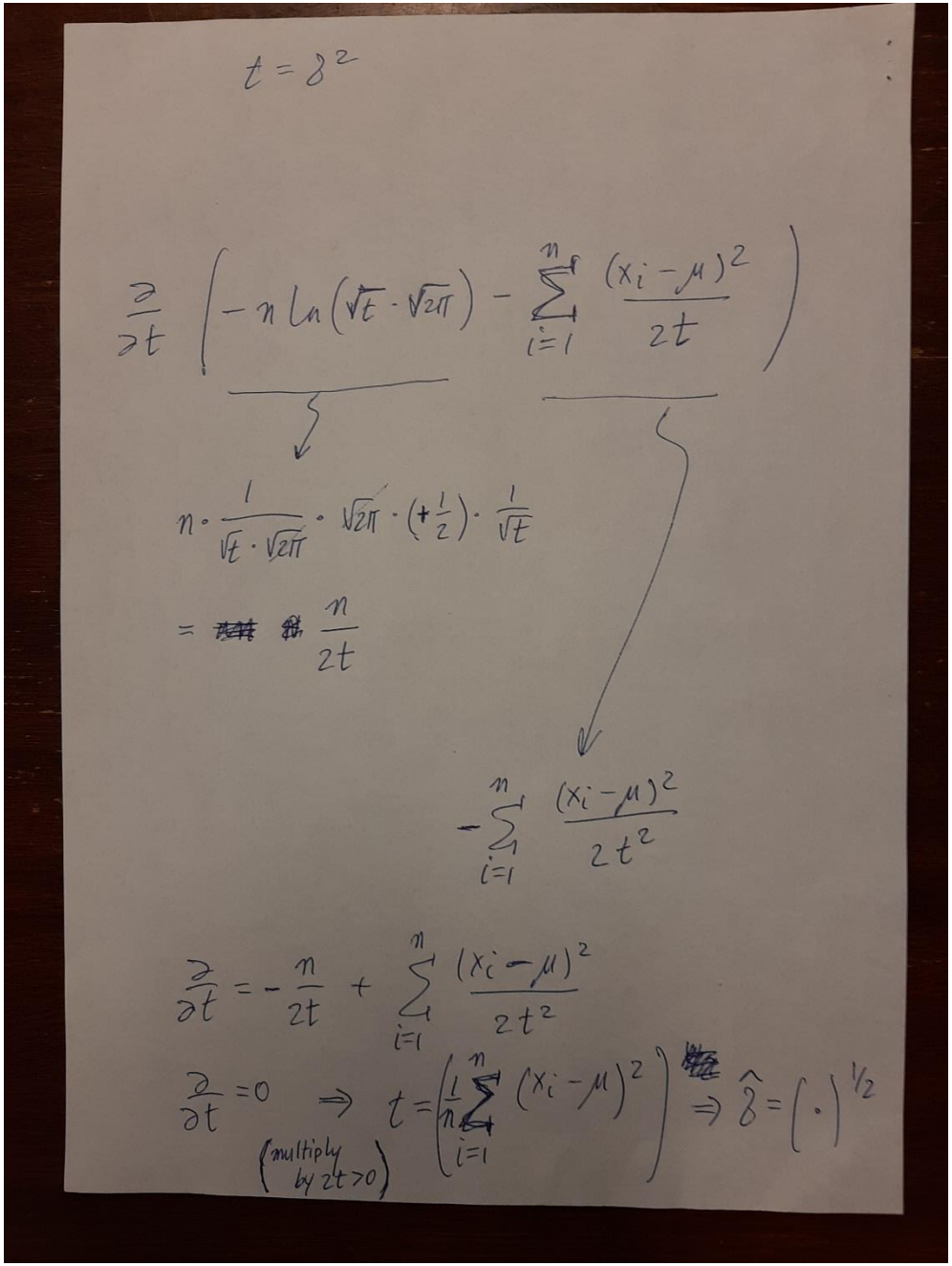
$$L(\theta) = -n\ln\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

To find a max point for this function we set the gradient to 0:

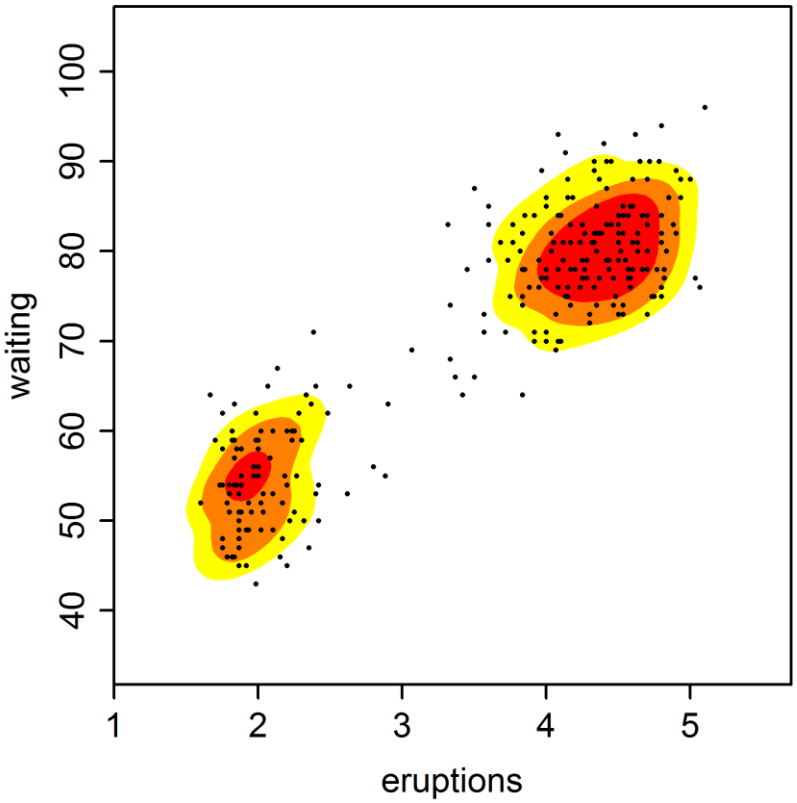
$$\frac{\partial}{\partial\mu} : \sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial(\sigma^2)} : -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}(\mu) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{1/2}$$

Expansion of this last step ….

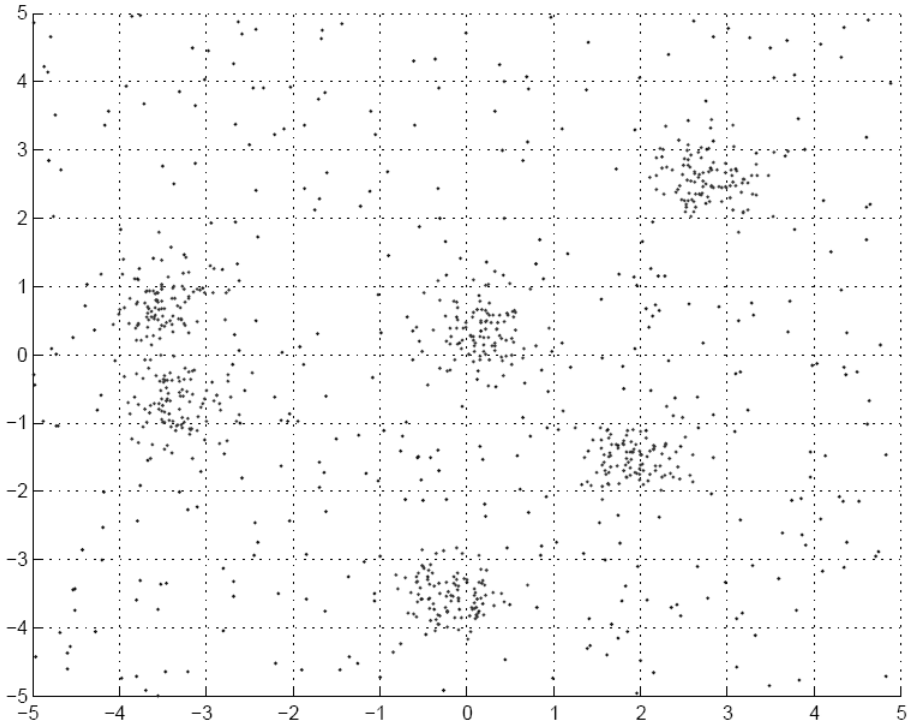


Old Faithful Wyoming



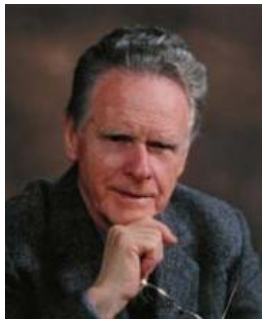
- What can we observe?
- What would we like to know in order to understand the entire system (and make predictions)?
- Gaussian Mixtures

Several underlying distributions



Expectation Maximization (**EM**) Algorithm

- Iterative method for parameter estimation where layers of data are missing from the observation
- Arthur Dempster, Nan Laird,
  Donald Rubin,
  J of the Royal Stat Soc, 1977
- Many variations followed. Research into methodology, applications and implementations is very active
- Has two steps:
  Expectation (E) and Maximization (M)
- Applicable to a wide range of machine learning and inference tasks



Harvard Univ, Dept of Statistics

MAP classification (next week)

Classes $A_1, \ldots, A_k$

Classify an instance with observed properties $\vec{x}$ as coming from the class indexed as

$$i^* = \underset{i}{\operatorname{argmax}}\, P(\vec{x}|A_i)P(A_i)$$

Summary

- Fitting distribution models to data is very useful for learning and other prediction/assessment/modelling tasks
- MLE approach
- Examples: Bernoulli, Univariate Gauss
- Next topic – Bayes classifiers
- EM – two weeks from now
- We will also further discuss multivariate Gaussians and GMMs