Probabilistic Perspective of Learning

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

Outline

- Introduction
- Parameter estimation
  - Maximum-Likelihood (ML) estimation (Frequentist approach)
  - Maximum A Posteriori (MAP) estimation (Bayesian approach)
- Probabilistic perspective on regression

- In the next lecture
  - Probabilistic classification

Relation of learning & statistics

- Target model in the learning problems can be considered as a statistical model

- For a fixed set of data and underlying target (statistical model), the estimation methods try to estimate the target from the available data

Density estimation

- Estimating the probability density function $p(\boldsymbol{x})$, given a set of data points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ drawn from it.

- Main approaches of density estimation:
  - Parametric: assuming a parameterized model for density function
    - A number of parameters are optimized by fitting the model to the data set

  - Nonparametric (Instance-based): No specific parametric model is assumed
    - The form of the density function is determined entirely by the data

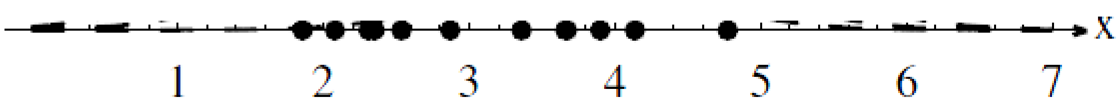
Parametric density estimation

- Goal: estimate parameters of a distribution from a dataset $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$
  - $\mathcal{D}$ contains N independent, identically distributed (i.i.d.) training samples.

- Assume that $p(\boldsymbol{x})$ in terms of a specific functional form which has a number of adjustable parameters.

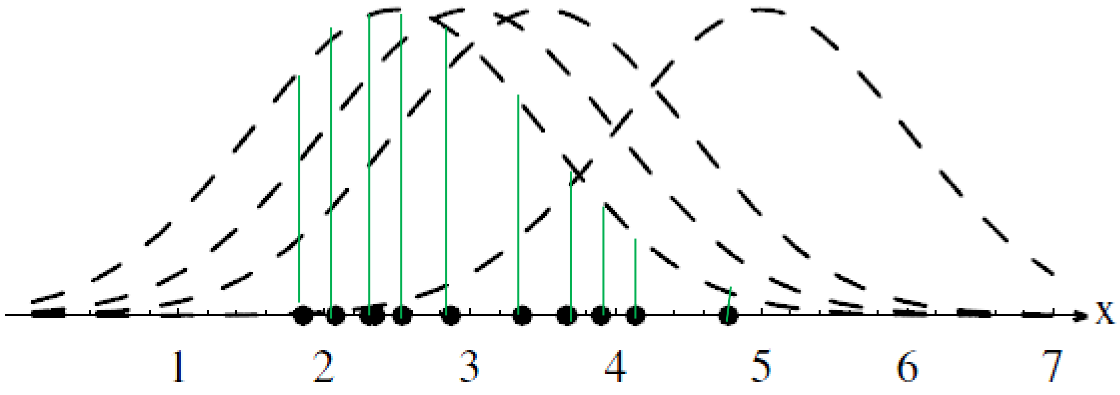
- We need to determine $\boldsymbol{\theta}$ given $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$
  - How to see $\boldsymbol{\theta}$?
    - A fixed and unknown number
    - A random variable

Example

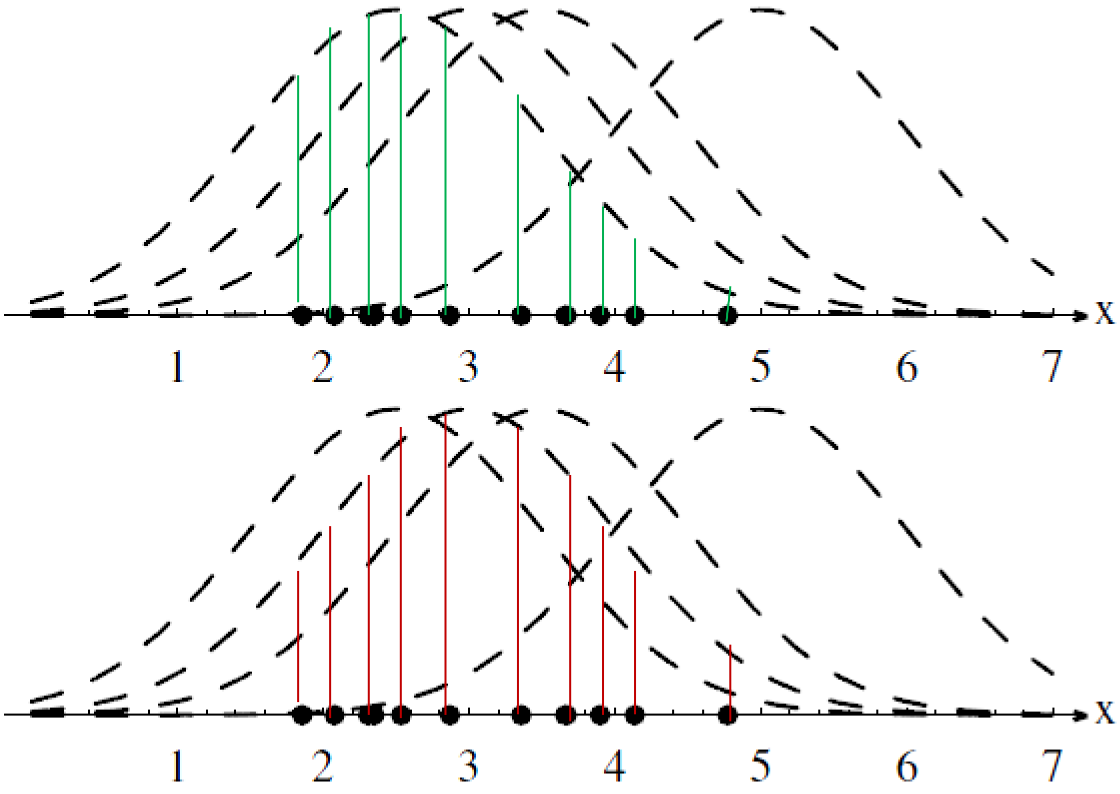


$$P(x|\mu) = N(x|\mu, 1)$$

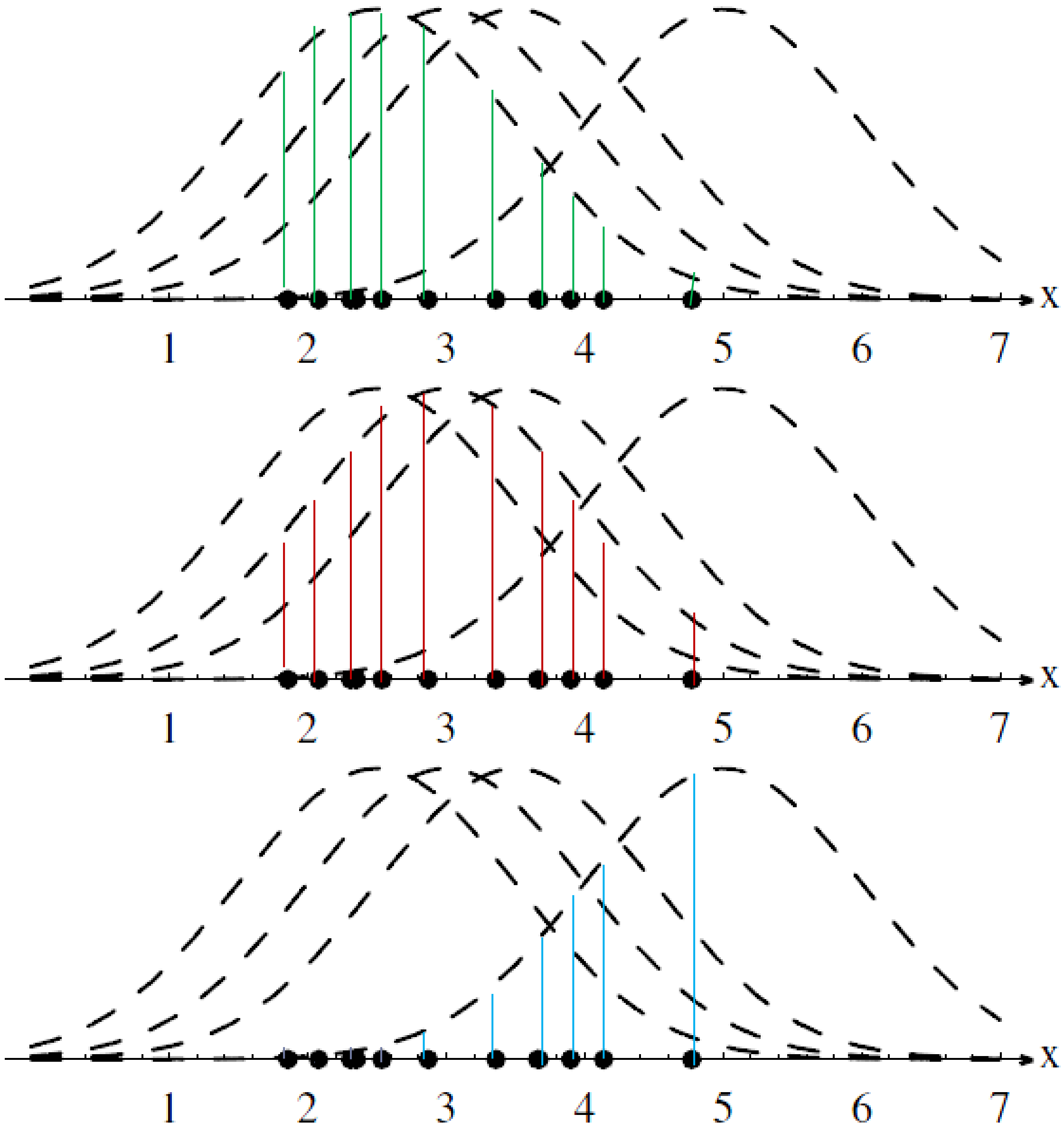
Example



Example



Example



Maximum Likelihood Estimation (MLE)

- A method of estimating the parameters of a statistical model given data.

- Likelihood is the conditional probability of observations $\mathcal{D} = \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}\right\}$ given the value of parameters $\boldsymbol{\theta}$
  - Assuming i.i.d. observations:

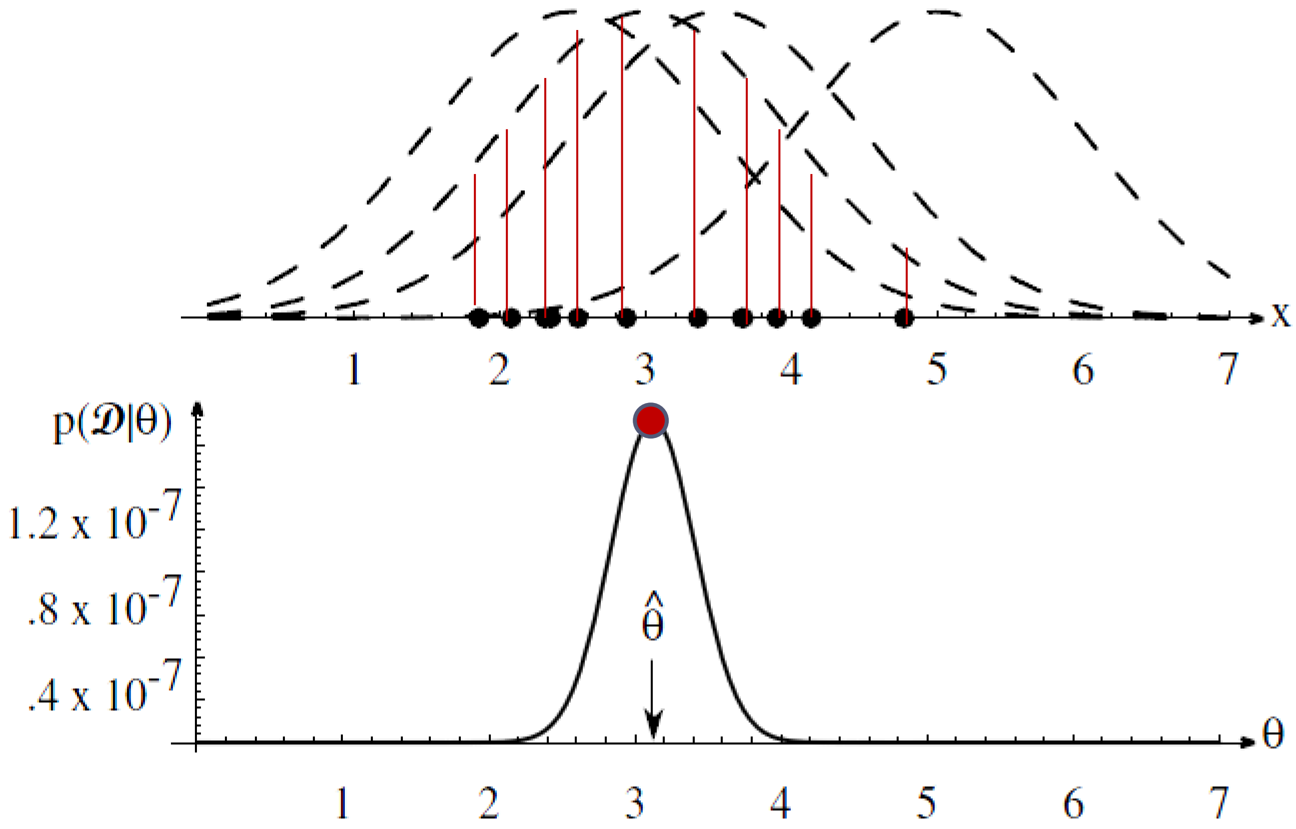
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

likelihood of $\boldsymbol{\theta}$ w.r.t. the samples

- Maximum Likelihood estimation

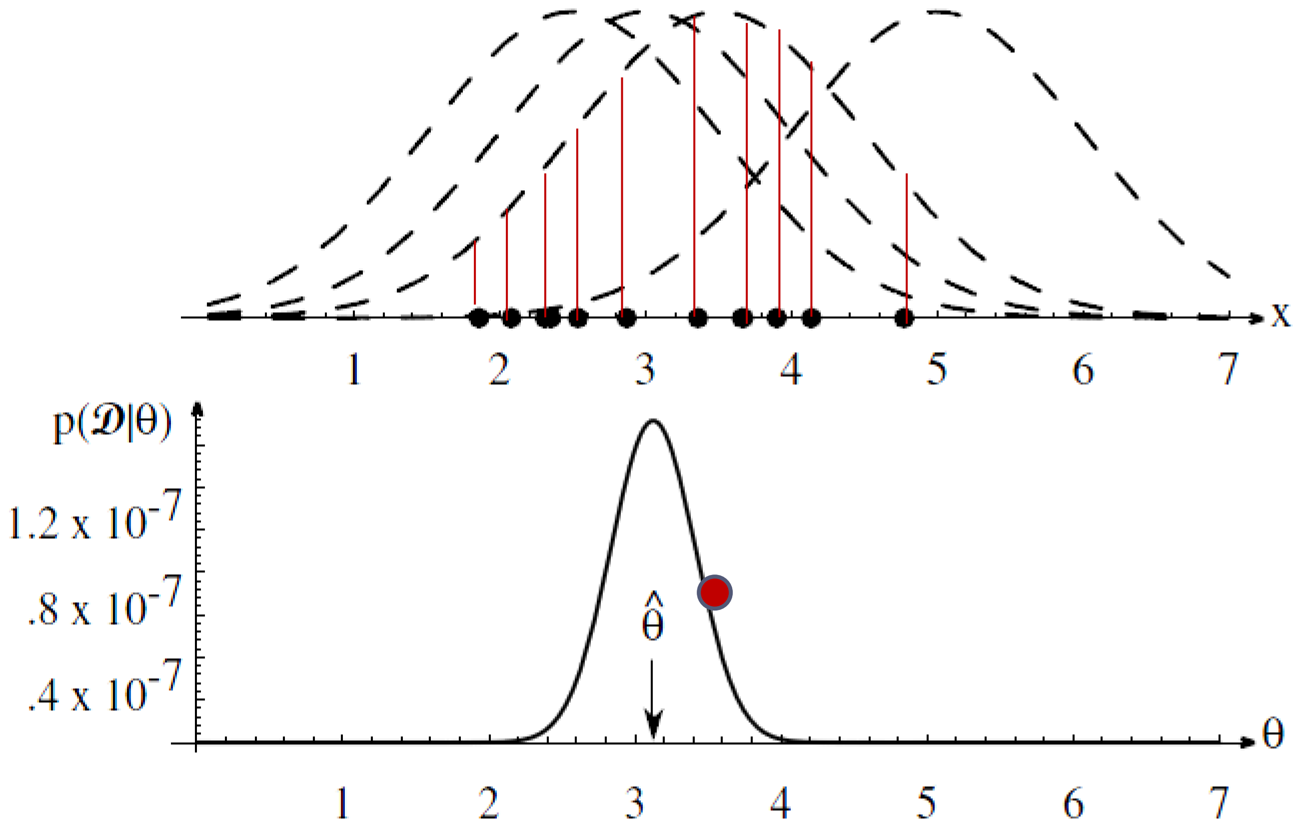
$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

Maximum Likelihood Estimation (MLE)



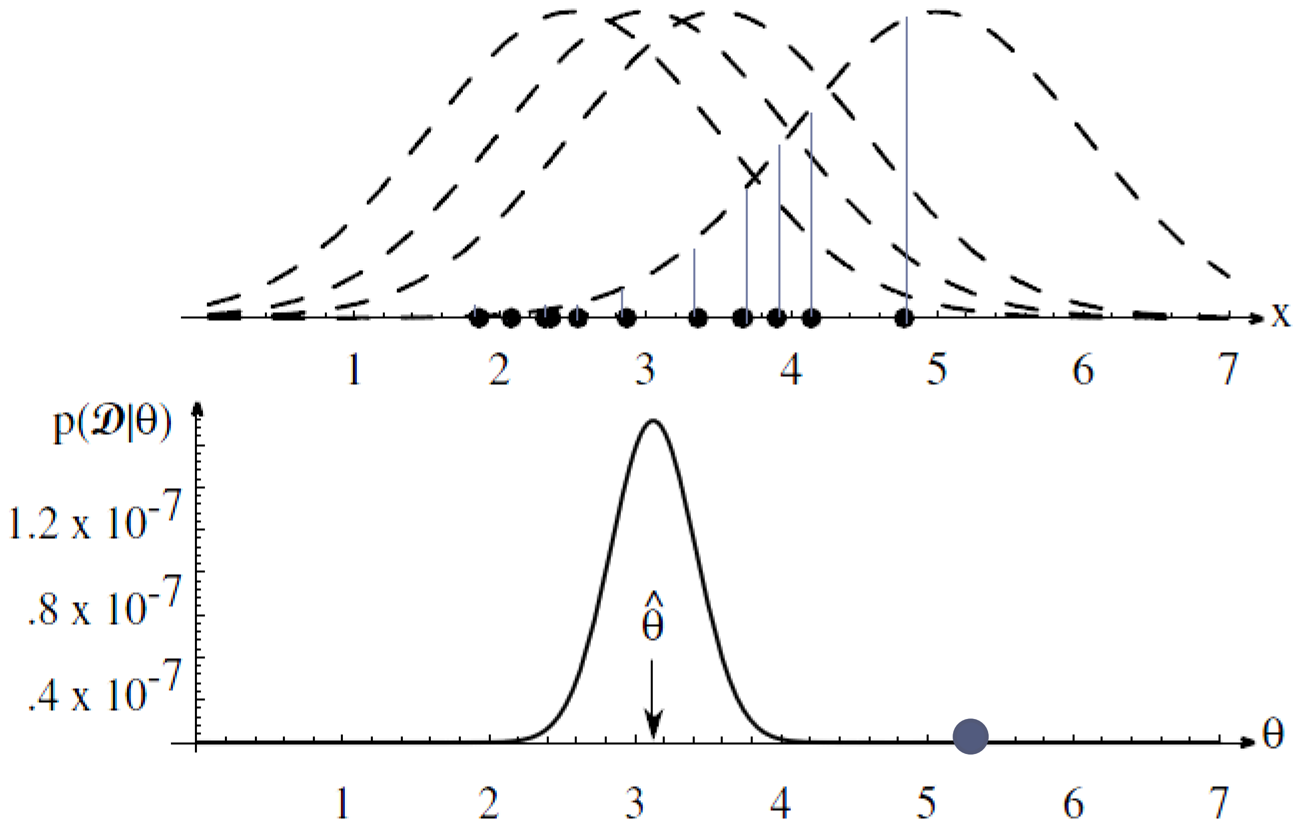
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)



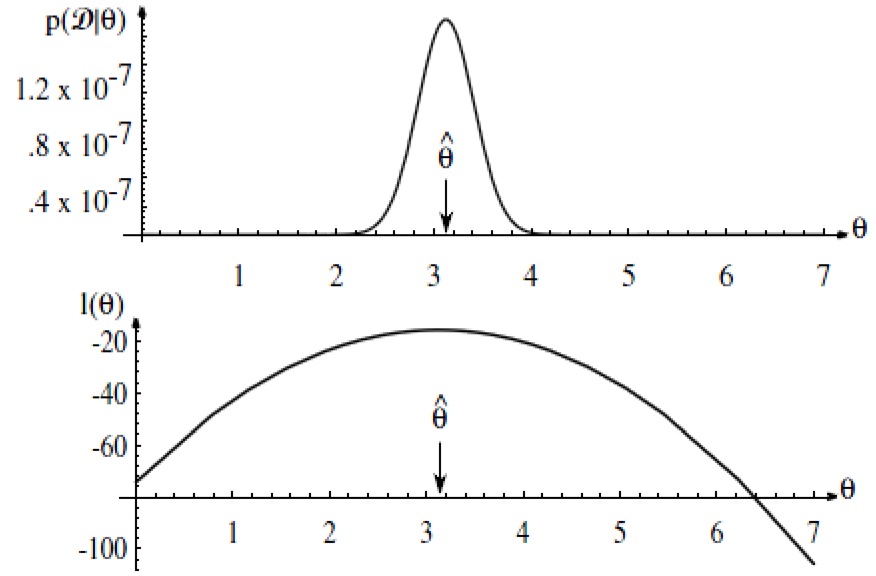
$\hat{\theta}$ best agrees with the observed samples

Maximum Likelihood Estimation (MLE)

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}\middle|\boldsymbol{\theta}\right)$$

- Thus, we solve $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$ to find global optimum



MLE: Bernoulli example

- A discrete example
- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$
  - m heads (1) - $N - m$ tails (0)
  - Bernoulli distribution

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

- The likelihood function

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{1-x^{(i)}}$$

MLE: Bernoulli example

- Maximizing the likelihood function

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta)$$

$$= \sum_{i=1}^{N} \{x^{(i)} \ln \theta + (1 - x^{(i)}) \ln(1 - \theta)\}$$

$$\frac{\partial \ln p(\mathcal{D}|\theta)}{\partial \theta} = 0 \Rightarrow \theta_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N} = \frac{m}{N}$$

MLE: Bernoulli example

- Example: $\mathcal{D} = \{1,1,1\}$
  - $\hat{\theta}_{ML} = \frac{3}{3} = 1$
  - Prediction: all future tosses will land heads up

- Over-fitting to $\mathcal{D}$

MLE: Gaussian example, unknown μ

- A continues example

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

$$\mathcal{L}(\mu) = \ln p(x^{(i)}|\mu) = -\ln\{\sqrt{2\pi}\sigma\} - \frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2$$

- Maximizing the likelihood function

$$\frac{\partial \mathcal{L}(\mu)}{\partial \mu} = 0 \Rightarrow \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{N} \ln p\left(x^{(i)}\middle|\mu\right)\right) = 0 \Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2}\left(x^{(i)} - \mu\right) = 0$$

$$\Rightarrow \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

Maximum A Posteriori (MAP) estimation

- Considering $\boldsymbol{\theta}$ as a random variable
- MAP estimation:
  - Assuming a prior distribution on $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta})$$

  - Maximizes the posterior distribution

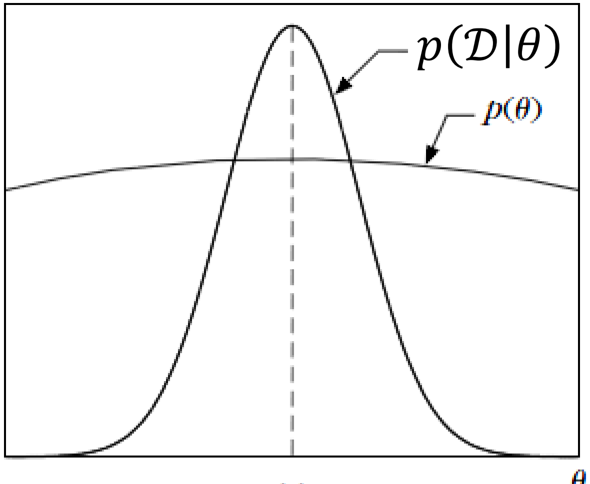
$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{D})$$

- Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

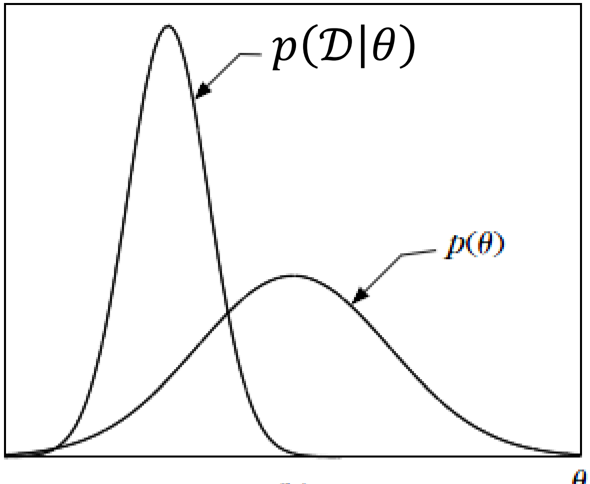
Maximum A Posteriori (MAP) estimation

- Given a set of observations $\mathcal{D}$ and a prior distribution $p(\boldsymbol{\theta})$ on parameters, the parameter vector that maximizes $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is found.



(a)

$\hat{\theta}_{MAP} \cong \hat{\theta}_{ML}$



(b)

$\hat{\theta}_{MAP} > \hat{\theta}_{ML}$

Example:
MAP: Bernoulli likelihood

- Given: $\mathcal{D} = \left\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\right\}$
  - m heads (1), $N - m$ tails (0)

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$= \left(\prod_{i=1}^{N} \theta^{x^{(i)}}(1-\theta)^{\left(1-x^{(i)}\right)}\right)\underbrace{\mathrm{Beta}(\theta|\alpha_1,\alpha_0)}_{\propto\, \theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}}$$

- Conjugate priors: The posterior distribution that is proportional to $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ will have <u>the same functional form</u> as the prior.

Example:
MAP: Bernoulli likelihood

- Given: $\mathcal{D} = \left\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\right\}$
  - m heads (1), $N - m$ tails (0)

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$= \left(\prod_{i=1}^{N} \theta^{x^{(i)}} (1-\theta)^{(1-x^{(i)})}\right) \underbrace{\mathrm{Beta}(\theta|\alpha_1, \alpha_0)}$$

$$\propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

$$\propto \theta^{m+\alpha_1-1}(1-\theta)^{N-m+\alpha_0-1}$$

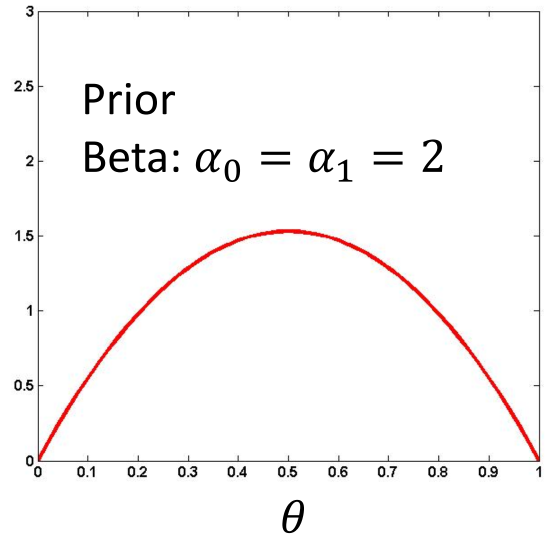
$$m = \sum_{i=1}^{N} x^{(i)}$$

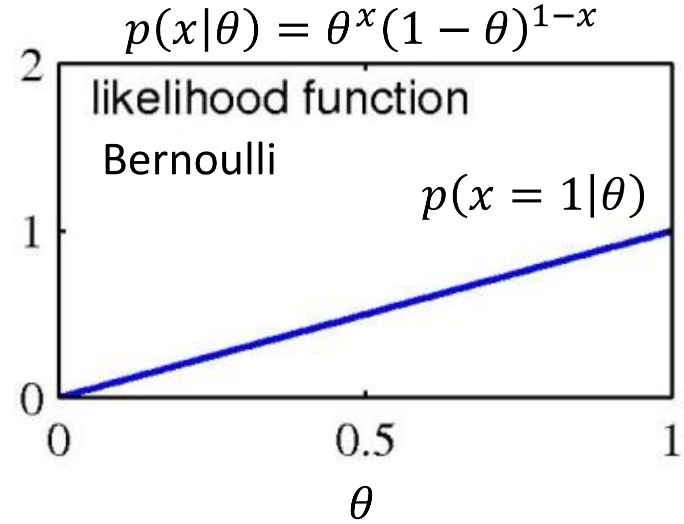
$$\Rightarrow p(\theta|\mathcal{D}) \propto Beta(\theta|\alpha_1', \alpha_0')$$

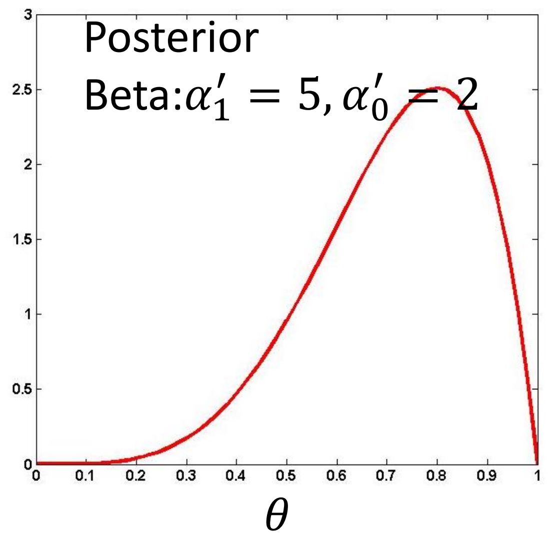
$$\alpha_1' = \alpha_1 + m$$

$$\alpha_0' = \alpha_0 + N - m$$

Example:
MAP: Bernoulli likelihood







Example:
MAP: Bernoulli likelihood

- Toss example: MAP estimation can avoid overfitting
  - $\mathcal{D} = \{1,1,1\}$, $\hat{\theta}_{ML} = 1$
  - $\hat{\theta}_{MAP} = 0.8$ (with prior $p(\theta) = \text{Beta}(\theta|2,2)$)

$$\alpha_0 = \alpha_1 = 2$$

$$\mathcal{D} = \{1,1,1\} \Rightarrow \; N = 3, m = 3$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|\mathcal{D}) = \frac{\alpha_1' - 1}{\alpha_1' - 1 + \alpha_0' - 1} = \frac{4}{5}$$

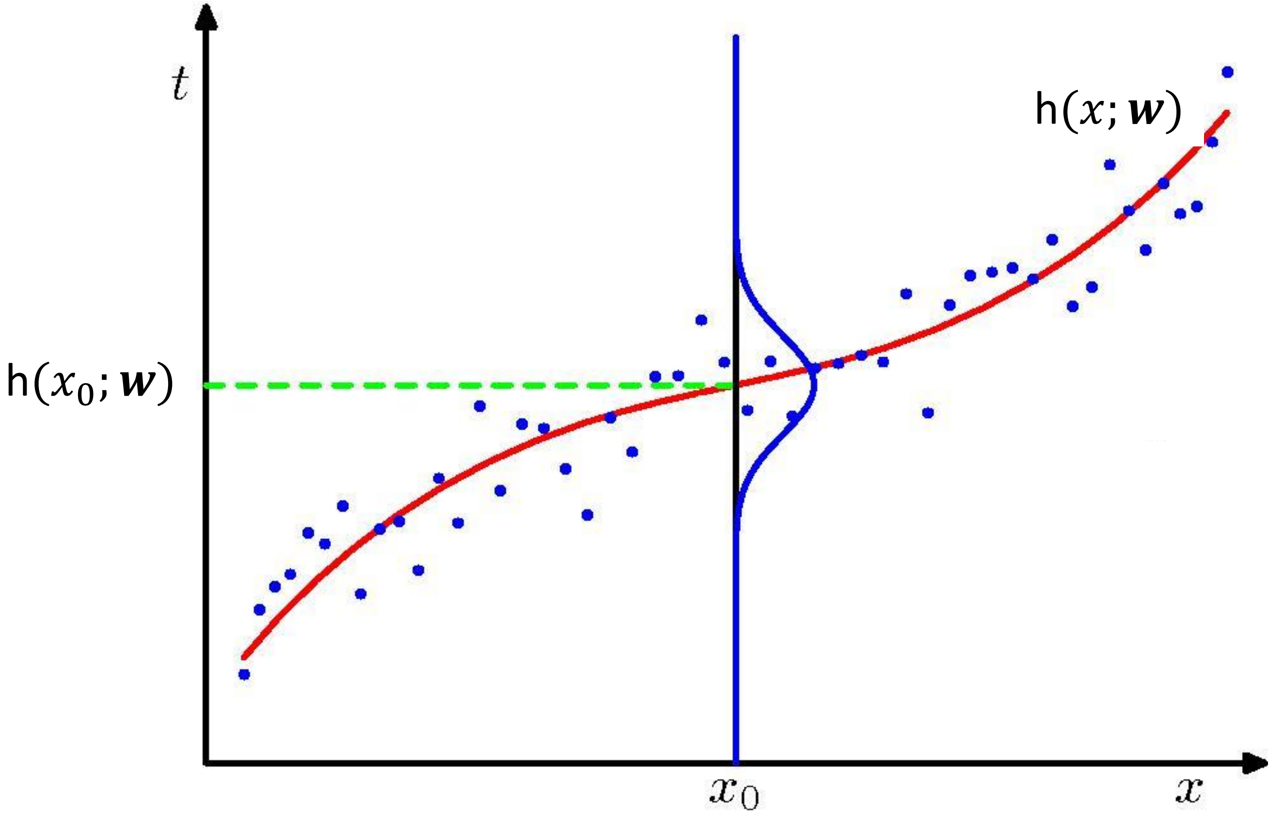
Summary

- ML and MAP result in a single (point) estimate of the unknown parameters vector.
  - More simple and interpretable
  - Alternative: Bayes estimator

- Two methods asymptotically ($N \to \infty$) results in the same estimate.

Probabilistic perspective on regression

- Describing uncertainty over value of target variable as a probability distribution



Probabilistic perspective on regression Example

- Special case:

Observed output = function + noise

$$y = h(\boldsymbol{x}; \boldsymbol{w}) + \epsilon$$

$$\text{e.g., } \epsilon \sim N(0, \sigma^2)$$

- The distribution of output, conditioned on the input variable:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = N(h(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$$

- Noise: Whatever we cannot capture with our chosen family of functions

Probabilistic perspective on regression
Example

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$
- Find the parameters that maximize the (conditional) likelihood of the outputs:

$$L(\mathcal{D}; \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \cdots & x_d^{(n)} \end{bmatrix}$$

Probabilistic perspective on regression
Example

- Univariate regression
  - Considering

$$h(x) = w_0 + w_1 x$$

  - We have

$$p(y|x, w, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(y - w_0 - w_1 x)^2)$$

- MLE for parameter estimation

Probabilistic perspective on regression Example

- Maximize the likelihood of the outputs (i.i.d):

$$L(\mathcal{D};\boldsymbol{w},\sigma^2) = \prod_{i=1}^{N} p(y^{(i)}|\boldsymbol{x}^{(i)},\boldsymbol{w},\sigma^2)$$

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, L(\mathcal{D};\boldsymbol{w},\sigma^2) = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{i=1}^{N} p(y^{(i)}|\boldsymbol{x}^{(i)},\boldsymbol{w},\sigma^2)$$

- It is often easier (but equivalent) to try to maximize the <u>log-likelihood</u>:

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ln p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\sigma^2)$$

Probabilistic perspective on regression
Example

- Maximize the log-likelihood:

$$\ln \prod_{i=1}^{N} p(y^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{w}, \sigma^2) = \sum_{i=1}^{N} \ln \mathcal{N}(y^{(i)} | f(\boldsymbol{x}^{(i)}; \boldsymbol{w}), \sigma^2)$$

$$= -N \ln \sigma - \frac{N}{2} \ln 2\pi - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{N} \left(y^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{w})\right)^2}_{\text{sum of squares error}}$$

- Maximizing log-likelihood (when we assume $y = \mathrm{h}(\boldsymbol{x}; \boldsymbol{w}) + \epsilon, \epsilon \sim N(0, \sigma^2)$) is equivalent to minimizing SSE

Probabilistic perspective on regression
Example

- MAP Estimation
  - Given observations $\mathcal{D}$
  - Find the parameters that maximize the probabilities of the parameters after observing the data (posterior probabilities):

$$\boldsymbol{\theta}_{MAP} = \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}))$$

- Since $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_{MAP} = \max_{\boldsymbol{\theta}} \; p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Probabilistic perspective on regression
Example

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$

$$\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w})$$

- The prior distribution on parameters

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \alpha^2 \boldsymbol{I}) = \left(\frac{1}{\sqrt{2\pi}\alpha}\right)^{d+1} exp\left\{-\frac{1}{2\alpha^2}\boldsymbol{w}^T\boldsymbol{w}\right\}$$

Probabilistic perspective on regression Example

- Given observations $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$

$$\max_{\boldsymbol{w}} \ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w})$$

$$\min_{\boldsymbol{w}} \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(y^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{w})\right)^2 + \frac{1}{\alpha^2} \boldsymbol{w}^T \boldsymbol{w}$$

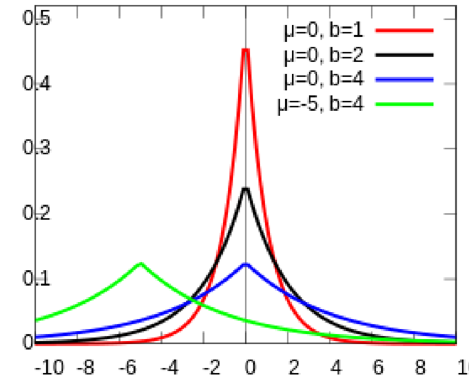
- Equivalent to regularized SSE with $\lambda = \frac{\sigma^2}{\alpha^2}$

Probabilistic perspective on regression
Example

- The prior distribution on parameters
  - Laplace distribution

$$p(\boldsymbol{w}) = Laplace(\boldsymbol{0}, b) = \frac{1}{2b} exp\left\{-\frac{|\boldsymbol{w}|}{b}\right\}$$

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y^{(i)} - f(\boldsymbol{x}^{(i)}; \boldsymbol{w})\right)^2 + \lambda\, |\boldsymbol{w}|$$



- Equivalent to the well known Lasso form for sparse regression

References

- [1] Mahdieh Soleymani, Machine learning, Sharif university of technology
- [2] C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 2.