Fundamentals of Machine Learning

MODEL FITTING & PARAMETER ESTIMATION

LEAST SQUARES

Amit K Roy-Chowdhury

Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee

Adapted from slides of Prof. Fowler @ Binghamton University;
Modified by Prof. Roy-Chowdhury @ UC Riverside

Model Fitting / Training

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \boxed{\mathcal{L}(\boldsymbol{\theta})}$$

Loss function / objective function

Maximum likelihood estimation

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Estimated Parameter

Probability

Model

Since most optimization algorithms are designed to minimize cost functions, we redefine the objective function to be the (conditional) negative log likelihood or NLL and we minimize NLL

$$\text{NLL}(\boldsymbol{\theta}) \triangleq -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Notation and Form for MAP

Notation: $\hat{\theta}_{MAP}$ maximizes the posterior PDF

$$\hat{\theta}_{MAP} = \arg\max_{\theta}\; p(\theta \,|\, \mathbf{x})$$

Equivalent Form (via Bayes' Rule): $\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta}\, [\, p(\mathbf{x} \,|\, \theta)\; p(\theta)]}$

Proof: Use $p(\theta \,|\, \mathbf{x}) = \dfrac{p(\mathbf{x} \,|\, \theta)\, p(\theta)}{p(\mathbf{x})}$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left[\frac{p(\mathbf{x} \,|\, \theta)\, p(\theta)}{\boxed{p(\mathbf{x})}} \right] = \arg\max_{\theta}\, [\, p(\mathbf{x} \,|\, \theta)\; p(\theta)]$$

Least Squares Approach

All the previous methods we've studied… required a probabilistic model for the data: Needed the PDF $p(\mathbf{x};\theta)$

For a Signal + Noise problem we needed:

Signal Model & Noise Model

Least-Squares is not statistically based!!!

⇒ Do NOT need a PDF Model

Linear Least Squares

A <u>linear</u> least-squares problem is one where the parameter observation model is linear: $\underbrace{\mathbf{s}}_{N\times 1} = \underbrace{\mathbf{H}}_{N\times p \text{ Known Matrix}}\underbrace{\boldsymbol{\theta}}_{p\times 1}$ $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e}$

***p* = <u>Order</u> of the model**

We must assume that **<u>H</u>** <u>is full rank</u>… otherwise there are multiple parameter vectors that will map to the same **s!!!**

Linear Least Squares

For the linear model the LS cost is: $J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} \left(x[n] - s[n;\boldsymbol{\theta}] \right)^2$

$$= \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta} \right)^T \left(\mathbf{x} - \mathbf{H}\boldsymbol{\theta} \right)$$

Now, to minimize, first expand:

$$J(\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} \boxed{- \mathbf{x}^T\mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x}} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}$$

$$= \mathbf{x}^T\mathbf{x} \boxed{- 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta}} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}$$

Scalar = scalar^T So…
$\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x} = (\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x})^T = \mathbf{x}^T\mathbf{H}\boldsymbol{\theta}$

Now setting $\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ gives $-2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{0}$

$$\Rightarrow \boxed{\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}^T\mathbf{x}}$$

Called the
"LS Normal Equations"

Because **H** is full rank we know that $\mathbf{H}^T\mathbf{H}$ is invertible:

$$\Rightarrow \boxed{\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}} \Rightarrow \boxed{\hat{\mathbf{s}}_{LS} = \mathbf{H}\hat{\boldsymbol{\theta}}_{LS} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}}$$

Comparing LSE to MLE

Model

Estimate

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{e}$$

No Probability Model Needed

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

PDF Gaussian, White

$$\hat{\theta}_{ML} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

If you assume Gaussian & apply this… BUT you are WRONG … you at least get the LSE!

Empirical Risk Minimization

We can generalize MLE by replacing the (conditional) log loss term, with any other loss function.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n)$$

This is known as empirical risk minimization or ERM, since it is the expected loss where the expectation is taken wrt the empirical distribution.