Lecture 10: Linear Regression

Introduction to Machine Learning [25737]

Sajjad Amini

Sharif University of Technology

Contents

References

Except explicitly cited, the reference for the material in slides is:

- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT press.

Section 1

Approach Definitions

Approach Definitions

Regression

- Task T: Finding mapping $f : \boldsymbol{x} \mapsto \boldsymbol{y}$ ($\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^D$ and $y \in \mathbb{R}$)
- Experience E: Set of N input-output pairs $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
- $P = \frac{1}{N}\sum_{n=1}^N (y_n - f(\boldsymbol{x}_n; \boldsymbol{\theta}))^2$

Linear Regression

Similar to classification problems, in regression problems we model $p(y|\boldsymbol{x}, \boldsymbol{\theta})$. Linear regression is the class of regression problem modeling where the expected value of the output is assumed to be a linear function of the input. In other words:

$$\mathbb{E}[y|\boldsymbol{x}, \boldsymbol{\theta}] = \boldsymbol{w}^T \boldsymbol{x}$$

where $\boldsymbol{w}$ is a subset of model parameters $\boldsymbol{\theta}$.

Section 2

Linear Regression Model

Linear Regression Model

Linear Regression Model

One model for linear regression can be formulated as:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|w_0 + \boldsymbol{w}^T\boldsymbol{x}, \sigma^2)$$

where:

$\boldsymbol{w}$	Weights or regression coefficients
w_0	Bias or offset
σ^2	Estimation variance

and $\boldsymbol{\theta} = [w_0; \boldsymbol{w}; \sigma^2]$.

Vectors Augmentation

Similar to classification models, we usually consider augmented vectors $[w_0; \boldsymbol{w}]$ and $[1; \boldsymbol{x}]$, which results in the following model:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x}, \sigma^2)$$

In this case $\boldsymbol{\theta} = [\boldsymbol{w}; \sigma^2]$.

Linear Regression Model Extensions

Extension to Vector Response $\boldsymbol{y}$

Consider the situation where response is vector $\boldsymbol{y} \in \mathbb{R}^J$ rather than scalar. Then assuming the elements of $\boldsymbol{y}$ are independent, we have:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y_j|\boldsymbol{w}_j^T \boldsymbol{x}, \sigma_j^2)$$

where $\boldsymbol{\theta} = [\boldsymbol{w}_1; \ldots; \boldsymbol{w}_J; \sigma_1^2; \ldots; \sigma_J^2]$.

Feature Transformation

Similar to classification problems, we can use feature transformation to reach a more descriptive models as:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma^2)$$

where $\boldsymbol{\theta} = [\boldsymbol{w}; \sigma^2]$.

MLE

Using model formulation, we have:

$$p(\mathcal{D}|\boldsymbol{\theta}) \overset{(1)}{=} p(\{y_n\}_{n=1}^N | \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{\theta}) \overset{(2)}{=} \prod_{i=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta})$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right)$$

where we use mode definition and independence of training samples in equality (1) and (2), respectively. Thus negative log-likelihood is:

$$\mathrm{NLL}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right)\right]$$

$$= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \widehat{y}_n)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

where we define $\widehat{y}_n \triangleq \boldsymbol{w}^T\boldsymbol{x}_n$.

MLE

Converting Summation into Matrix Form

We can easily show that:

$$\mathrm{RSS}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2 = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = \frac{1}{2}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

where RSS$(\cdot)$ stand for *residual sum of squares* function and:

$$\boldsymbol{X} = \begin{bmatrix} - & \boldsymbol{x}_1 & - \\ & \vdots & \\ - & \boldsymbol{x}_n & - \end{bmatrix}, \ \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Thus the negative log-likelihood can be written as:

$$\mathrm{NLL}(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

MLE

MLE

The optimal value for $\boldsymbol{w}$ and σ^2 can be calculated as:

$$\nabla_{\boldsymbol{w}} \mathrm{NLL}(\boldsymbol{\theta}) = \mathbf{0} \Rightarrow \nabla_{\boldsymbol{w}} \mathrm{RSS}(\boldsymbol{w}) = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{y} = \mathbf{0}$$

$$\Rightarrow \widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$\nabla_{\sigma} \mathrm{NLL}(\boldsymbol{\theta}) = 0 \Rightarrow \widehat{\sigma}^2_{mle} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \widehat{\boldsymbol{w}}^T_{mle} \boldsymbol{x}_n)^2 = \frac{2}{N} \mathrm{RSS}(\widehat{\boldsymbol{w}}_{mle})$$

Note that $\widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ is known as *Ordinary Least Squares* (OLS) solution.

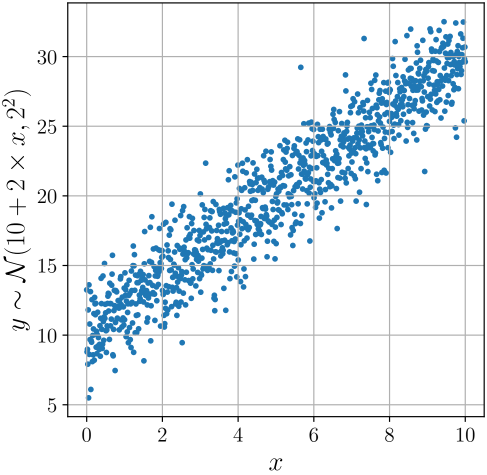
Sample: Modeling a Device Measurement

Data Generation

Assume we have device that generates its measurements based on the following rule:

$$\begin{cases} y_i \sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i = 10 + 2 \times x_i \\ \sigma^2 = 2^2 \end{cases}$$

On the right, we see a realization of this data for $N = 1000$.



Sample: Modeling a Device Measurement

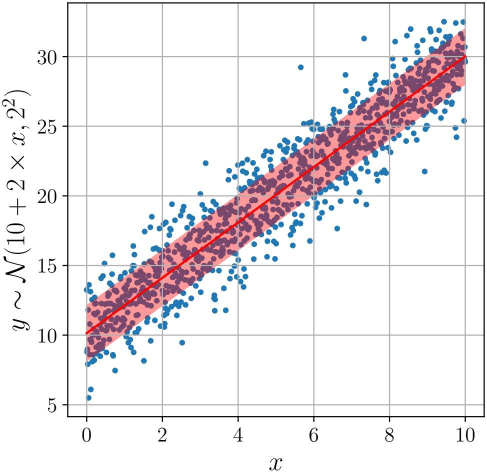
Data Generation

Solution: Based on the MLE fol-rmulation, the solution is:

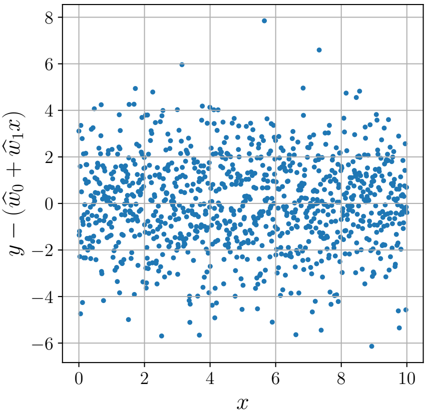
$$\widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = [10.14, 1.99]^T$$

$$\widehat{\sigma}^2_{mle} = \frac{1}{1000}\sum_{n=1}^{1000}(y_n - \widehat{\boldsymbol{w}}^T_{mle}\boldsymbol{x}_n)^2 = 3.85$$

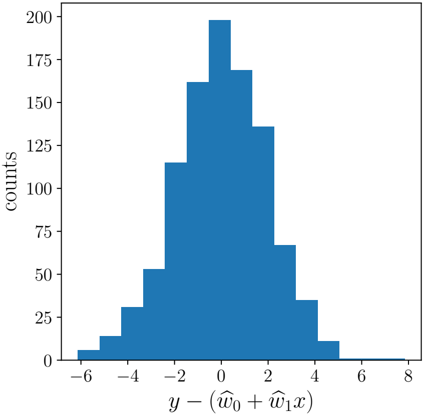
On the right, we see the solution.



Sample: Modeling a Device Measurement



(a) Difference



(b) Difference Histogram

Figure: Visualization of difference

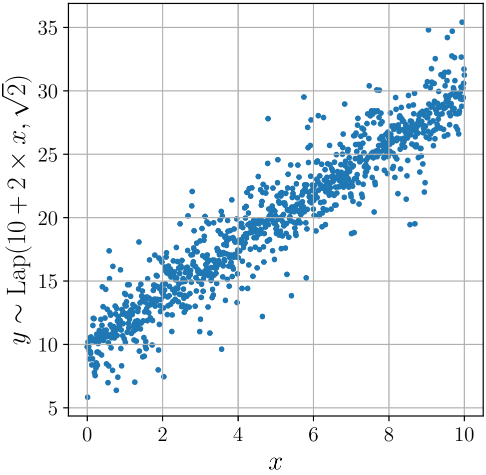
Sample: Modeling a Device Measurement

Data Generation

Assume we have device that generates its measurements based on the following rule:

$$\begin{cases} y_i \sim \text{Lap}(\mu_i, b) \\ \mu_i = 10 + 2 \times x_i \\ b = \sqrt{2} \end{cases}$$

On the right, we see a realization of this data for $N = 1000$.



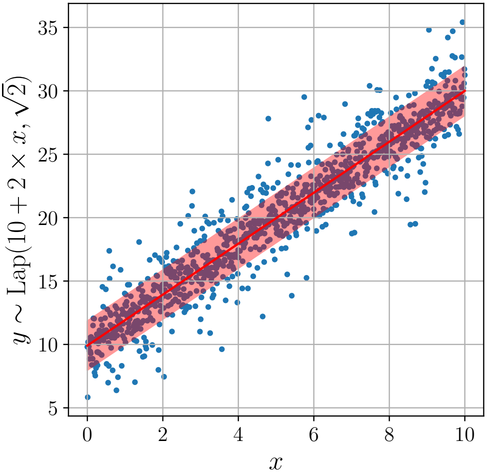
Data Generation

Solution: Based on the MLE folrmulation, the solution is:

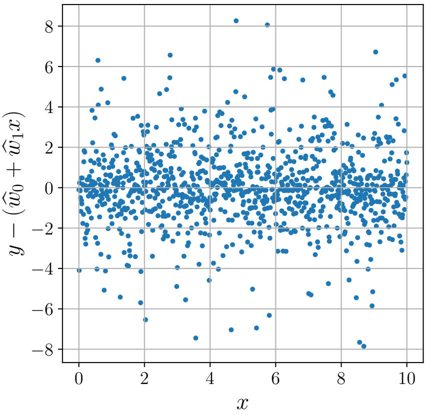
$$\widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = [9.91, 2.01]^T$$

$$\widehat{\sigma}^2_{mle} = \frac{1}{1000}\sum_{n=1}^{1000}(y_n - \widehat{\boldsymbol{w}}^T_{mle}\boldsymbol{x}_n)^2 = 3.74$$

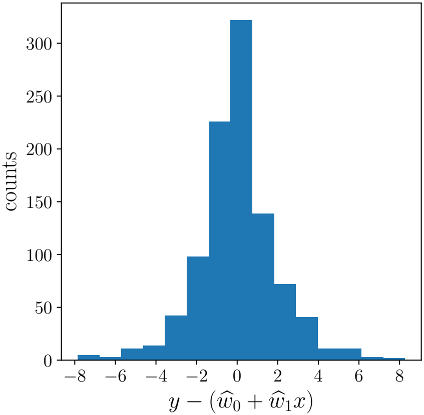
On the right, we see the solution.



Sample: Modeling a Device Measurement



(a) Difference



(b) Difference Histogram

Figure: Visualization of difference

Section 3

Heteroskedastic Regression (Weighted LR)

Heteroskedastic Regression(Weighted LR)

Heteroskedastic Regression

Heteroskedastic Regression assume the follwoing model for data:

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x}, \sigma^2(\boldsymbol{x})) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^T\boldsymbol{x})^2\right)$$

MLE

For the simplicity, assume we have access to $\{\sigma^2(\boldsymbol{x}_n)\}_{n=1}^N$, then:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x}_n)}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x}_n)}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right)$$

$$\mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \boldsymbol{\Lambda}^{-1}),\ \boldsymbol{\Lambda} = \mathrm{diag}(\frac{1}{\sigma^2(\boldsymbol{x}_n)})$$

Heteroskedastic Regression (Weighted LR)

MLE

$$p(\mathcal{D}|\boldsymbol{\theta}) = \frac{1}{|2\pi\boldsymbol{\Lambda}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\right)$$

$$\Rightarrow \mathrm{NLL}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T \boldsymbol{\Lambda}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \mathrm{NLL}(\boldsymbol{\theta}) = \boldsymbol{0} \Rightarrow \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} = \boldsymbol{0} \Rightarrow \widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y}$$

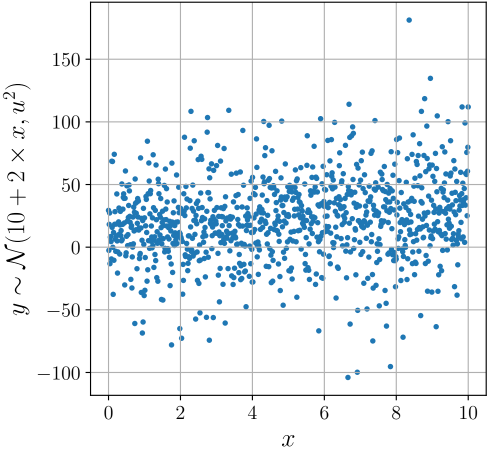
The above is known as *Weighted Least Squares* (WLS) estimate.

Data Generation

Assume we have device that generates its measurements based on the following rule:

$$\begin{cases} y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i = \beta_0 + \beta_1 x_i \\ \sigma_i^2 = u_i^2 \\ u_i \propto U(10, 50) \end{cases}$$

On the right, we see a realization of this data for $N = 1000$.



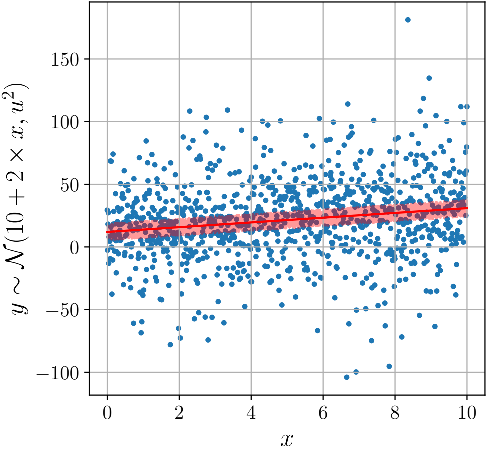
Sample: Modeling a Device Measurement

MLE Solution

Solution: Based on OLS, we have:

$$\widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$= [11.97, 1.90]^T$$

Below we see the solution.



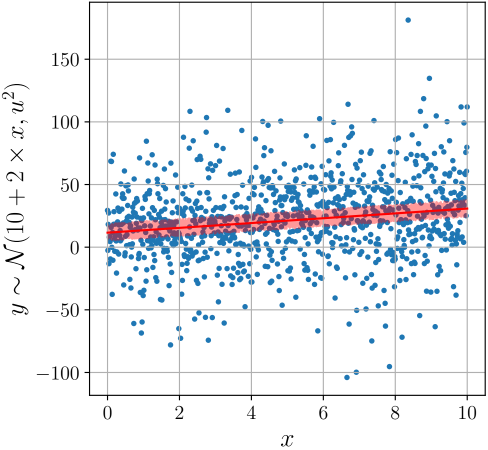
Sample: Modeling a Device Measurement

MLE Solution

Solution: Based on WLS, we have:

$$\begin{aligned}\widehat{\boldsymbol{w}}_{mle} &= (\boldsymbol{X}^T\boldsymbol{\Lambda}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}\boldsymbol{y} \\ &= [11.57, 1.92]^T\end{aligned}$$

Below we see the solution.

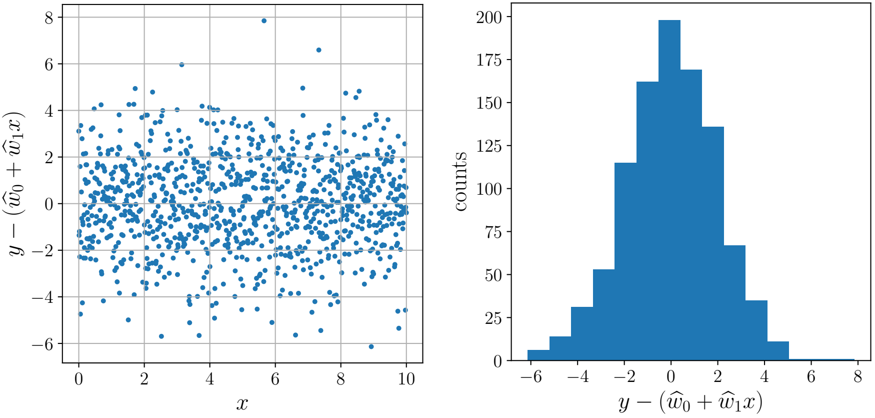


Section 4

Measuring Goodness of Fit

Residual Plots

For one dimensional inputs, we can plot the residual $r_n = y_n - \widehat{y}_n$ vs the input x_n. The resulting plot is called residual plot. This plot should be similar to samples of $\mathcal{N}(0, \sigma^2)$.



Coefficient of Determination

Coefficient of Determination

Coefficient of Determination is defined as:

$$R^2 \triangleq 1 - \frac{\sum_{n=1}^{N} (\widehat{y}_n - y_n)^2}{\sum_{n=1}^{N} (\bar{y} - y_n)^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n$ and $\text{TSS} = \sum_{n=1}^{N} (\bar{y}_n - y_n)^2$ stands for total sum of squares (TSS). You can show that $0 \leq R^2 \leq 1$

Section 5

MAP Estimation (Regularization)

Subsection 1

Ridge Regression

Ridge Regression

Ridge Regression

To avoid overfitting in linear regression, similar to classification models, we can assume the weight vector to come from the following prior:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \tau^2 \boldsymbol{I})$$

where λ is a hyper-parameter. The resulting MAP estimation is known as Ridge Regression and is formulated as:

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

Ridge Regression

MAP

Assuming the σ to be known, we have:

$$
\begin{aligned}
p(\boldsymbol{w}|\mathcal{D}) \propto & \overbrace{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right)}^{p(\mathcal{D}|\boldsymbol{w})} \overbrace{\frac{1}{|2\pi\tau^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T(\tau^2\boldsymbol{I})^{-1}\boldsymbol{w}\right)}^{p(\boldsymbol{w})} \\
\propto & \frac{1}{|2\pi\sigma^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T\right) \\
& \frac{1}{|2\pi\tau^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T(\tau^2\boldsymbol{I})^{-1}\boldsymbol{w}\right)
\end{aligned}
$$

Thus we have:

$$
\begin{aligned}
\underset{\boldsymbol{w}}{\operatorname{argmax}}\ p(\boldsymbol{w}|\mathcal{D}) \equiv \underset{\boldsymbol{w}}{\operatorname{argmin}}\ -\log p(\boldsymbol{w}|\mathcal{D}) = \underset{\boldsymbol{w}}{\operatorname{argmin}}\ & \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{2\tau^2}\|\boldsymbol{w}\|^2 \\
\equiv & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \overbrace{\frac{\sigma^2}{\tau^2}}^{\lambda}\|\boldsymbol{w}\|^2
\end{aligned}
$$

Ridge Regression

MAP

$$
\begin{aligned}
& J(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2 \\
\Rightarrow & \nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y} + \lambda\boldsymbol{w}) = \mathbf{0} \\
\Rightarrow & \widehat{\boldsymbol{w}}_{map} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \left(\sum_n \boldsymbol{x}_n\boldsymbol{x}_n^T + \lambda\boldsymbol{I}\right)^{-1}\left(\sum_n y_n\boldsymbol{x}_n\right)
\end{aligned}
$$

Subsection 2

Lasso Regression

Feature Selection

Feature Selection

Assume the problem we encountered in ridge regression:

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

Now change the ℓ_2 norm with ℓ_0 norm as:

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_0}_{J_0(\boldsymbol{w})}$$

where $\|\boldsymbol{w}\|_0 = \sum_{d=1}^{D} \mathbb{I}(|w_d| > 0)$. In the above formulation, the problem is solved by using reduced number of features.

Sample: Modeling a Device Measurement

Data Generation

Assume we have device that generates its measurements based on $\begin{cases} y_i \sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i = 10 + 2 \times x_i \\ \sigma^2 = 2^2 \end{cases}$

Then we have the following states:

- $\beta_0 \neq 0, \beta_1 \neq 0 \Rightarrow J(\widetilde{\boldsymbol{w}}) = 3848.2 + \lambda \times 2$
- $\beta_0 \neq 0, \beta_1 = 0 \Rightarrow J(\widetilde{\boldsymbol{w}}) = 36842.5 + \lambda \times 1$
- $\beta_0 = 0, \beta_1 \neq 0 \Rightarrow J(\widetilde{\boldsymbol{w}}) = 29584.4 + \lambda \times 1$
- $\beta_0 = 0, \beta_1 = 0 \Rightarrow J(\widetilde{\boldsymbol{w}}) = 439953.5 + \lambda \times 0$

Thus we have the following result for the $J_0(\boldsymbol{w})$ problem:

$$\widehat{\boldsymbol{w}} = \begin{cases} [0,0]^T \ (J(\widehat{\boldsymbol{w}}) = 439953.5) & 439953.5 - 29584.4 < \lambda \\ [0,3.51]^T \ (J(\widehat{\boldsymbol{w}}) = 29584.4 + \lambda \times 1) & 29584.4 - 3848.2 < \lambda < 439953.5 - 2958 \\ [10.13, 1.98]^T \ (J(\widehat{\boldsymbol{w}}) = 3848.2 + \lambda \times 2) & \lambda < 29584.4 - 3848.2 \end{cases}$$

Sample: Modeling a Device Measurement

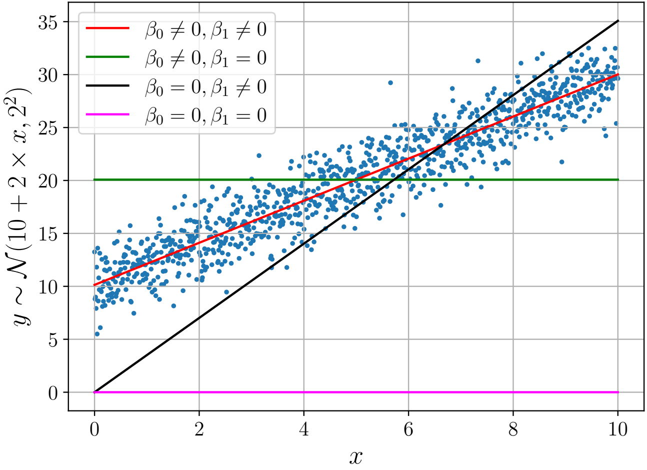


Figure: Realization of device data for $N = 1000$

Lasso Regression

Lasso Regression

Assume dimensions of weight vector are independently and identically distributed as $\text{Lap}(w|0, b)$. Then the prior distribution over weight vector is:

$$p(\boldsymbol{w}) = \prod_{d=1}^{D} \frac{1}{2b} \exp\left(-\frac{|w_i|}{b}\right)$$

where b is a hyper-parameter. The resulting MAP estimation is known as Lasso (Least Absolute Shrinkage and Selection Operator) Regression and is formulated as:

$$\widehat{\boldsymbol{w}}_{map} = \underset{\boldsymbol{w}}{\text{argmax}}\, p(\boldsymbol{w}|\mathcal{D})p(\boldsymbol{w})$$

Ridge Regression

MAP

Assuming the σ to be known, we have:

$$p(\boldsymbol{w}|\mathcal{D}) \propto \overbrace{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right)}^{p(\mathcal{D}|\boldsymbol{w})} \overbrace{\frac{1}{(2b)^D} \exp\left(-\frac{1}{b}\sum_{d=1}^{D}|w_i|\right)}^{p(\boldsymbol{w})}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T\right) \exp\left(-\frac{1}{b}\|\boldsymbol{w}\|_1\right)$$

Thus we have:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}}\ p(\boldsymbol{w}|\mathcal{D}) \equiv \underset{\boldsymbol{w}}{\operatorname{argmin}}\ -\log p(\boldsymbol{w}|\mathcal{D}) = \underset{\boldsymbol{w}}{\operatorname{argmin}}\ \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{b}\|\boldsymbol{w}\|_1$$

$$\equiv \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{\overbrace{2\sigma^2}^{\lambda}}{b}\|\boldsymbol{w}\|_1$$

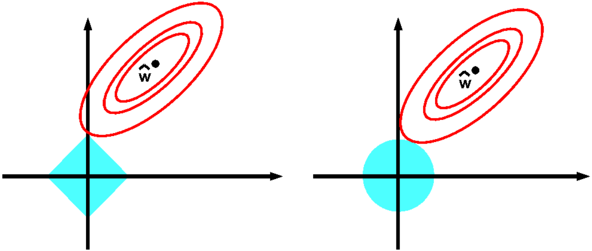
Connection to Feature Selection

Lagrangian Interpretation

Using Lagrangian interpretation, we have:

$$\text{RidgeRegression} : \min_{\boldsymbol{w}} \ \text{NLL}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2 \equiv \begin{cases} \min_{\boldsymbol{w}} \ \ \text{NLL}(\boldsymbol{w}) \\ \text{s.t.} \ \|\boldsymbol{w}\|_2^2 \leq B \end{cases}$$

$$\text{LassoRegression} : \min_{\boldsymbol{w}} \ \text{NLL}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 \equiv \begin{cases} \min_{\boldsymbol{w}} \ \ \text{NLL}(\boldsymbol{w}) \\ \text{s.t.} \ \|\boldsymbol{w}\|_1 \leq C \end{cases}$$



MAP

For the solution of Lasso regression, we only consider the case where $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. In this case, we have:

$$\widehat{\boldsymbol{w}}_{mle} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}$$

If we use the above, we have:

$$\begin{aligned} J(\boldsymbol{w}) =& \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1 \\ =& \boldsymbol{y}^T\boldsymbol{y} - 2\overbrace{\boldsymbol{y}^T\boldsymbol{X}}^{\widehat{\boldsymbol{w}}_{mle}^T}\boldsymbol{w} + \boldsymbol{w}^T\overbrace{\boldsymbol{X}^T\boldsymbol{X}}^{\boldsymbol{I}}\boldsymbol{w} + \lambda\|\boldsymbol{w}\|_1 + \widehat{\boldsymbol{w}}_{mle}^T\widehat{\boldsymbol{w}}_{mle} - \overbrace{\widehat{\boldsymbol{w}}_{mle}^T\widehat{\boldsymbol{w}}_{mle}}^{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y}} \\ =& \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{mle}\|^2 + \lambda\|\boldsymbol{w}\|_1 + \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{y} \end{aligned}$$

Thus we have:

$$\begin{aligned} \min_{\boldsymbol{w}} \ J(\boldsymbol{w}) &\equiv \min_{\boldsymbol{w}} \ \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{mle}\|^2 + \lambda\|\boldsymbol{w}\|_1 \\ &\equiv \min_{w_i} \ (w_i - \widehat{w}_{(mle)i})^2 + \lambda|w_i|, \ i = 1, \ldots, D \end{aligned}$$

Lasso Regression

MAP

We should solve the problem of the following form:

$$\widehat{w}_{lasso} = \underset{w}{\operatorname{argmin}} \ \overbrace{\left(w - \widehat{w}_{mle}\right)^2 + \lambda|w|}^{J(w)}$$

Now assume two cases:

- $\widehat{w}_{mle} > 0 \Rightarrow w > 0$, then:

$$J(w) = \left(w - \widehat{w}_{mle}\right)^2 + \lambda w \Rightarrow \frac{d}{d\,w} J(w) = 2(w - \widehat{w}_{mle}) + \lambda = 0 \Rightarrow w = \widehat{w}_{mle} - \frac{\lambda}{2}$$

- $\widehat{w}_{mle} \leq 0 \Rightarrow w \leq 0$, then:

$$J(w) = \left(w - \widehat{w}_{mle}\right)^2 - \lambda w \Rightarrow \frac{d}{d\,w} J(w) = 2(w - \widehat{w}_{mle}) - \lambda = 0 \Rightarrow w = \widehat{w}_{mle} + \frac{\lambda}{2}$$

Altogether we have: $\widehat{w}_{lasso} = \begin{cases} \max\{\widehat{w}_{mle} - \frac{\lambda}{2}, 0\} & \widehat{w}_{mle} > 0 \\ \min\{\widehat{w}_{mle} + \frac{\lambda}{2}, 0\} & \widehat{w}_{mle} \leq 0 \end{cases} = \mathcal{S}(\widehat{w}_{mle}, \frac{\lambda}{2})$

Soft Thresholding Operator

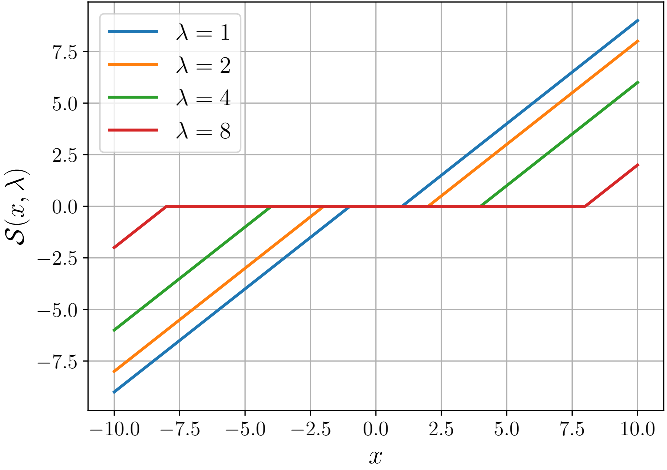


Figure: Soft Thresholding Operator Curve

Section 6

Bayesian Linear Regression

Bayesian Linear Regression

Prior and Likelihood

Assume the following prior distribution:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\breve{\boldsymbol{w}}, \breve{\boldsymbol{\Sigma}})$$

On the other hand, we see before that the likelihood can be written as (we assume σ^2 to be known):

$$p(\mathcal{D}|\boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{w}^T\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

Bayes Rule for Gaussian

If $\begin{cases} p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \\ p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y) \end{cases}$, then:

$$p(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}), \begin{cases} \boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W} \\ \boldsymbol{\mu}_{z|y}\boldsymbol{\Sigma}_{z|y}\left[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\mu}_z\right] \end{cases}$$

Posterior

Thus the posterior can be calculated as:

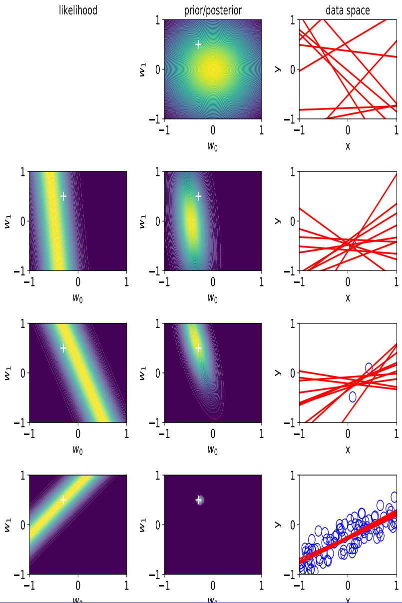
$$p(\boldsymbol{w}|\mathcal{D}) \propto \mathcal{N}(\boldsymbol{w}|\breve{\boldsymbol{w}}, \breve{\boldsymbol{\Sigma}})\mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}) = \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, \hat{\boldsymbol{\Sigma}})$$

where we have:

$$\hat{\boldsymbol{w}} \triangleq \hat{\boldsymbol{\Sigma}}(\breve{\boldsymbol{\Sigma}}^{-1}\breve{\boldsymbol{w}} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y})$$

$$\hat{\boldsymbol{\Sigma}} \triangleq (\breve{\boldsymbol{\Sigma}}^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Bayesian Linear Regression



Posterior Predictive

Normalization Constant in Bayes' Rule for Gaussian

In Bayes' rule for Gaussian, we have:

$$p(\boldsymbol{z}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{y})}$$

where the normalization factor is:

$$p(\boldsymbol{y}) = \int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)\mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y)d\boldsymbol{z} = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b}, \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T)$$

Computing Posterior Prediction

$$p(y|\boldsymbol{x}, \mathcal{D}) = \int \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})\mathcal{N}(y|\boldsymbol{x}^T\boldsymbol{w}, \sigma^2)d\boldsymbol{w} = \mathcal{N}(y|\hat{\boldsymbol{\mu}}^T\boldsymbol{x}, \hat{\sigma}^2(\boldsymbol{x}))$$

$$\hat{\sigma}^2(\boldsymbol{x}) = \sigma^2 + \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{x}$$