Regression and Regularization: Part II

Prof. Gustavo Nonato

NYU / CUSP - GX 5006

February 14, 2017

Linear Regression via Maximum Likelihood

Lets consider the linear model

$$y_i = \sum_{j=0}^{d} x_{ij}\beta_j + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id}) \in \mathbb{R}^{d+1}, y_i \in \mathbb{R},\ i = 1, \ldots, n.$

Linear Regression via Maximum Likelihood

Lets consider the linear model

$$y_i = \sum_{j=0}^{d} x_{ij}\beta_j + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id}) \in \mathbb{R}^{d+1}, y_i \in \mathbb{R},\ i = 1, \ldots, n.$
Assuming y_i i.i.d. and σ^2 known, we have:

$$p(y_i|\mathbf{x_i}, \boldsymbol{\beta}) \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

Linear Regression via Maximum Likelihood

Lets consider the linear model

$$y_i = \sum_{j=0}^{d} x_{ij}\beta_j + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id}) \in \mathbb{R}^{d+1}, y_i \in \mathbb{R},\ i = 1, \ldots, n.$
Assuming y_i i.i.d. and σ^2 known, we have:

$$p(y_i | \mathbf{x_i}, \boldsymbol{\beta}) \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

The likelihood function for y_i is given by:

$$p(D|\boldsymbol{\beta}) \ = p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i | \mathbf{x_i}, \boldsymbol{\beta})$$

Linear Regression via Maximum Likelihood

Lets consider the linear model

$$y_i = \sum_{j=0}^{d} x_{ij}\beta_j + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id}) \in \mathbb{R}^{d+1}, y_i \in \mathbb{R},\ i = 1, \ldots, n.$
Assuming y_i i.i.d. and σ^2 known, we have:

$$p(y_i|\mathbf{x_i}, \boldsymbol{\beta}) \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

The likelihood function for y_i is given by:

$$\underbrace{\boxed{p(D|\boldsymbol{\beta})}}_{\textit{likelihood}} = p(y_1, \ldots, y_n|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x_i}, \boldsymbol{\beta})$$

Linear Regression via Maximum Likelihood

$$p(D|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x_i}, \boldsymbol{\beta})$$

Linear Regression via Maximum Likelihood

$$p(D|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x_i}, \boldsymbol{\beta})$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2)$$

Linear Regression via Maximum Likelihood

$$p(D|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x_i}, \boldsymbol{\beta})$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2)$$

$$= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right)$$

Linear Regression via Maximum Likelihood

$$p(D|\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x_i}, \boldsymbol{\beta})$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2)$$

$$= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}\boxed{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}\right)$$

$$= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})\right)$$

Linear Regression via Maximum Likelihood

We are looking for parameters $\boldsymbol{\beta}$ that maximize the likelihood

$$p(D|\boldsymbol{\beta}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)$$

Linear Regression via Maximum Likelihood

We are looking for parameters $\boldsymbol{\beta}$ that maximize the likelihood

$$p(D|\boldsymbol{\beta}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)$$

However, maximizing $p(D|\boldsymbol{\beta})$ is equivalent to minimize

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})$$

Linear Regression via Maximum Likelihood

Differentiating $\mathcal{L}$ with respect to $\boldsymbol{\beta}$

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}$$

Linear Regression via Maximum Likelihood

Differentiating $\mathcal{L}$ with respect to $\boldsymbol{\beta}$

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}$$

Making $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = 0$ we get

$$\boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Linear Regression via Maximum Likelihood

Differentiating $\mathcal{L}$ with respect to $\boldsymbol{\beta}$

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}^\top \boldsymbol{\beta})$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}$$

Making $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = 0$ we get

$$\boxed{\boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}}$$

This is exactly the same solution given by the least-squares method.

Bayesian Linear Regression

Likelihood

$$p(D|\boldsymbol{\beta}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)$$

Bayesian Linear Regression

Likelihood

$$p(D|\boldsymbol{\beta}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)$$

The Bayesian approach relies on the Bayes' rule to compute the *posterior* distribution

$$p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta})\; p(\boldsymbol{\beta})$$

Bayesian Linear Regression

Likelihood

$$p(D|\boldsymbol{\beta}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)$$

The Bayesian approach relies on the Bayes' rule to compute the *posterior* distribution

$$p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta})\boxed{p(\boldsymbol{\beta})}$$

All the trick is to play with the prior distribution of $p(\boldsymbol{\beta})$

Bayesian Linear Regression

Lets suppose

$$p(\boldsymbol{\beta}) = N(0, \frac{1}{\gamma})$$

Bayesian Linear Regression

Lets suppose

$$p(\boldsymbol{\beta}) = N(0, \frac{1}{\gamma})$$

$$p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

$$p(\boldsymbol{\beta}|D) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})\right)\exp\left(-2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)$$

$$p(\boldsymbol{\beta}|D) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}^\top\boldsymbol{\beta}) - 2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)$$

Bayesian Linear Regression

$$p(\boldsymbol{\beta}|D) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta}) - 2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)$$

Bayesian Linear Regression

$$p(\boldsymbol{\beta}|D) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta}) - 2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)$$

Some algebraic computation allow to rewrite the expression above as

$$p(\boldsymbol{\beta}|D) \propto N(\mu, \Lambda^{-1}) \tag{1}$$

where

$$\mu = \frac{1}{\sigma^2}(\sigma^2\mathbf{X}^\top\mathbf{X} + \gamma I)^{-1}\mathbf{X}^\top\mathbf{y}$$

$$\Lambda = \sigma^2\mathbf{X}^\top\mathbf{X} + \gamma I$$

Bayesian Linear Regression

$$p(\boldsymbol{\beta}|D) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}^\top\boldsymbol{\beta}) - 2\gamma\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)$$

Some algebraic computation allow to rewrite the expression above as

$$p(\boldsymbol{\beta}|D) \propto N(\mu, \Lambda^{-1}) \tag{1}$$

where

$$\mu = \boxed{\frac{1}{\sigma^2}(\sigma^2\mathbf{X}^\top\mathbf{X} + \gamma I)^{-1}\mathbf{X}^\top\mathbf{y}} \tag{2}$$

$$\Lambda = \sigma^2\mathbf{X}^\top\mathbf{X} + \gamma I$$

PS. I) Defining $\lambda = \frac{\sigma^2}{\gamma}$ Equation (2) is exactly the ridge regression solution!!

PS. II) μ could also be obtained from (1) by computing the maximum a posterior (MAP) value of $\boldsymbol{\beta}$.

Bayesian Linear Regression

$$\mu = \frac{1}{\sigma^2}(\sigma^2 \mathbf{X}^\top \mathbf{X} + \ \gamma I \ )^{-1} \mathbf{X}^\top \mathbf{y}$$

Bayesian Linear Regression

$$\mu = \frac{1}{\sigma^2}(\sigma^2\mathbf{X}^\top\mathbf{X} + \ \gamma I\ )^{-1}\mathbf{X}^\top\mathbf{y}$$

γ defines the variance of the prior distribution $p(\boldsymbol{\beta}) = N(0, \frac{1}{\gamma})$

Bayesian Linear Regression

$$\mu = \frac{1}{\sigma^2}(\sigma^2 \mathbf{X}^\top \mathbf{X} + \ \gamma I \ )^{-1} \mathbf{X}^\top \mathbf{y}$$

γ defines the variance of the prior distribution $p(\boldsymbol{\beta}) = N(0, \frac{1}{\gamma})$
Making $\gamma \to \infty$ (uninformative prior) pushes the solution towards non-regularized least-squares.

Bayesian Regression Linear Regression

$$p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta}) \boxed{p(\boldsymbol{\beta})}$$

Bayesian Regression Linear Regression

$$p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta}) \boxed{p(\boldsymbol{\beta})}$$

It can be shown that if the prior $p(\boldsymbol{\beta})$ is defined as a Laplace distribution then the MAP estimate of $\boldsymbol{\beta}$ is the lasso solution !!

Making Prediction

We have just seen that the maximum a posteriori (MAP) and regularized least-squares estimates are equivalent.

Making Prediction

We have just seen that the maximum a posteriori (MAP) and regularized least-squares estimates are equivalent.

From this result we are tempted to assume that the Bayesian framework is simply re-interpretation of classical methods.

Making Prediction

We have just seen that the maximum a posteriori (MAP) and regularized least-squares estimates are equivalent.

From this result we are tempted to assume that the Bayesian framework is simply re-interpretation of classical methods.

This is not the case!

Predictions can be accomplished by marginalizing with respect to the parameters, enabling a quite flexible mechanism (see Gelman, Andrew, et al. Bayesian data analysis. Chapman & Hall/CRC, 2014.)

Non-linear basis

Most of the developments presented so far is also valid if we replace the linear basis $\mathbf{x}$ for functions $\phi_j(\mathbf{x})$, that is,

$$\mathbf{y}_i \sim \sum_{j=0}^{m} \beta_j \phi_j(\mathbf{x}_i)$$

Non-linear basis

Most of the developments presented so far is also valid if we replace the linear basis $\mathbf{x}$ for functions $\phi_j(\mathbf{x})$, that is,

$$\mathbf{y}_i \sim \sum_{j=0}^{m} \beta_j \phi_j(\mathbf{x}_i)$$

In this case, matrix $\mathbf{X}$ is replaced to X_ϕ

$$\mathbf{X}_\phi = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_m(\mathbf{x}_1) \\ & \vdots & \\ \phi_0(\mathbf{x}_n) & \cdots & \phi_m(\mathbf{x}_n) \end{bmatrix}$$

Non-linear basis

Most of the developments presented so far is also valid if we replace the linear basis $\mathbf{x}$ for functions $\phi_j(\mathbf{x})$, that is,

$$\mathbf{y}_i \sim \sum_{j=0}^{m} \beta_j \phi_j(\mathbf{x}_i)$$

In this case, matrix $\mathbf{X}$ is replaced to X_ϕ

$$\mathbf{X}_\phi = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_m(\mathbf{x}_1) \\ & \vdots & \\ \phi_0(\mathbf{x}_n) & \cdots & \phi_m(\mathbf{x}_n) \end{bmatrix}$$

For instance, the ridge regression solution is simply:

$$\boldsymbol{\beta} = (\mathbf{X}_\phi^\top \mathbf{X}_\phi + \lambda I)^{-1} \mathbf{X}_\phi^\top \mathbf{y}$$

Non-linear basis

Typical basis functions are polynomial basis such as

$$\phi_0(\mathbf{x}_i) = 1, \phi_1(\mathbf{x}_i) = x_{i1}, \phi_2(\mathbf{x}_i) = x_{i1}^2, \ldots, \phi_m(\mathbf{x}_i) = x_{ip}^q$$

where q is a maximal degree for the basis.

Non-linear basis

Typical basis functions are polynomial basis such as

$$\phi_0(\mathbf{x}_i) = 1, \phi_1(\mathbf{x}_i) = x_{i1}, \phi_2(\mathbf{x}_i) = x_{i1}^2, \ldots, \phi_m(\mathbf{x}_i) = x_{ip}^q$$

where q is a maximal degree for the basis.
Radial basis functions are also a widely employed

$$\phi_j(\mathbf{x}_i) = \exp(-\|\mathbf{x}_i - \mathbf{c}_j\|^2 / r^2)$$

Non-linear basis

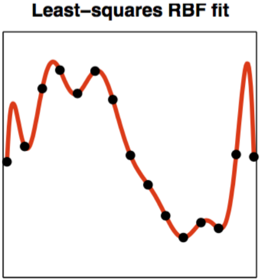
Typical basis functions are polynomial basis such as

$$\phi_0(\mathbf{x}_i) = 1, \phi_1(\mathbf{x}_i) = x_{i1}, \phi_2(\mathbf{x}_i) = x_{i1}^2, \ldots, \phi_m(\mathbf{x}_i) = x_{ip}^q$$

where q is a maximal degree for the basis.
Radial basis functions are also a widely employed

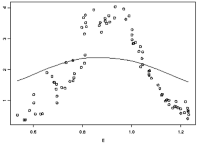
$$\phi_j(\mathbf{x}_i) = \exp(-\|\mathbf{x}_i - \mathbf{c}_j\|^2 / r^2)$$

The main advantage of using non-linear basis is that we can perform non-linear approximation using the linear regression framework.

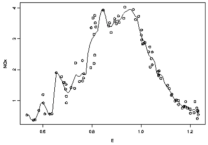


Bias-Variance Tradeoff

Model selection: estimating the performance of different models in order to choose the best one (for instance different values of λ in regularized LS).

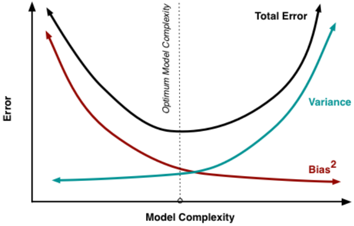


High Bias - Low Variance



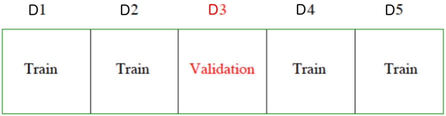
Low Bias - High Variance

“overfitting” - modeling the random component



Cross-Validation

K-fold approach:



D1	D2	D3	D4	D5
Train	Train	Validation	Train	Train

Given a model M and a K-fold of a data set D

- for $k = 1, \ldots, K$
  - Consider the training set $D^{(-k)} = D/D_k$
  - Learn M from $D^{(-k)}$
  - $e_k(M) = \sum_{i \in D_k} (y_i - \hat{y}_i^{(-k)})^2$
- $CV(M) = \frac{1}{n} \sum_{k=1}^{K} e_k(M)$

Cross-Validation

K-fold approach:

D1	D2	D3	D4	D5
Train	Train	Validation	Train	Train

Given a model M and a K-fold of a data set D

- for $k = 1, \ldots, K$
  - Consider the training set $D^{(-k)} = D/D_k$
  - Learn M from $D^{(-k)}$
  - $e_k(M) = \sum_{i \in D_k} (y_i - \hat{y}_i^{(-k)})^2$
- $CV(M) = \frac{1}{n} \sum_{k=1}^{K} e_k(M)$

When $K = n$ the K-fold is called *leave-one-out cross-validation*.

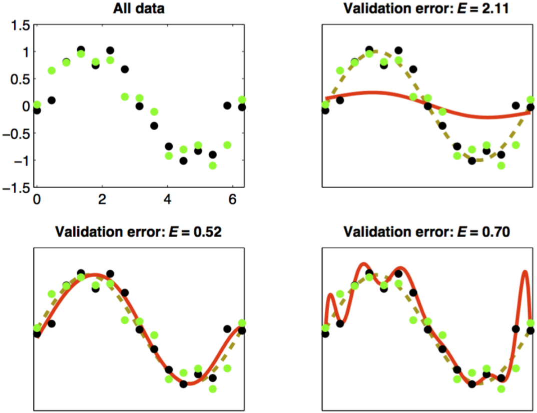
Cross-Validation

Model assessment: having chosen a model, estimating its prediction error on new data.

Cross-Validation

Model assessment: having chosen a model, estimating its prediction error on new data.

K-fold can be used to assess the quality of a particular model.



Recap

Summary of the lecture:

- Assuming Gaussian error, Maximum Likelihood Estimate is equivalent to non-regularized Least Squares;

Recap

Summary of the lecture:

- Assuming Gaussian error, Maximum Likelihood Estimate is equivalent to non-regularized Least Squares;
- With appropriate priors Maximum a Posteriori estimate is equivalent to Ridge and Lasso regression;

Recap

Summary of the lecture:

- Assuming Gaussian error, Maximum Likelihood Estimate is equivalent to non-regularized Least Squares;
- With appropriate priors Maximum a Posteriori estimate is equivalent to Ridge and Lasso regression;
- Non-linear basis functions allows for non-linear approximation using linear regression framework;

Recap

Summary of the lecture:

- Assuming Gaussian error, Maximum Likelihood Estimate is equivalent to non-regularized Least Squares;
- With appropriate priors Maximum a Posteriori estimate is equivalent to Ridge and Lasso regression;
- Non-linear basis functions allows for non-linear approximation using linear regression framework;
- Cross-validation is a fundamental tool to model selection and assessment.