Linear Regression and Gaussian Process

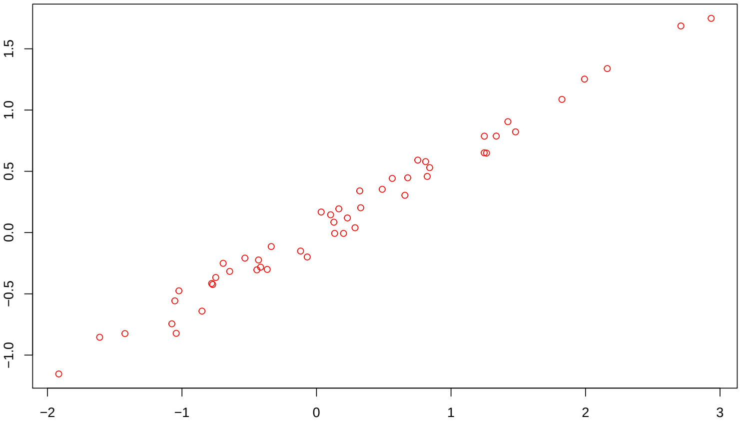
Tathagata Basu

UE de Master 2, AOS1

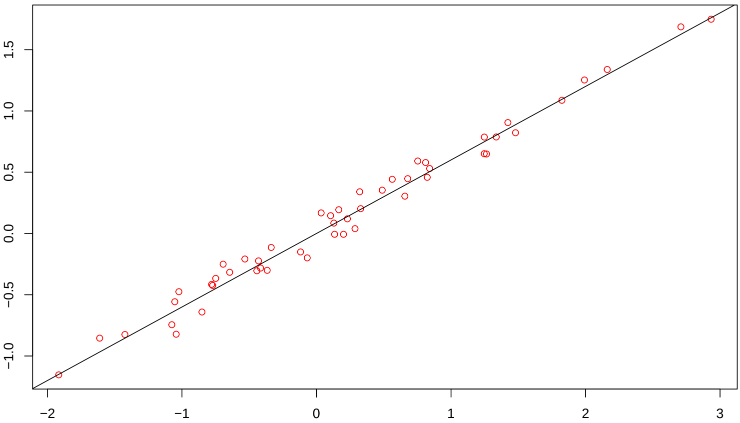
Autumn 2022

Outline I

Linear Model I



Linear Model II



Linear Model III

Let

$$Y = X\beta + \epsilon \tag{1}$$

where,

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{2}$$

$\beta = (\beta_1, \ldots, \beta_p)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, such that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Likelihood Function I

We know that $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ follows a normal distribution with mean 0 and variance σ^2. Therefore, the p.d.f is given by:

$$f(\epsilon_i \mid \beta_1, \cdots, \beta_p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) \tag{3}$$

Replacing ϵ_i with $y_i - \sum_{j=1}^{p} x_{ij}\beta_j$, we get

$$f(y_i, x_i \mid \beta_1, \cdots, \beta_p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2}\right) \tag{4}$$

Likelihood Function II

Then the likelihood function $\mathcal{L}(\beta \mid Y, X)$ is given by:

$$\mathcal{L}(\beta \mid Y, X) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \prod_{i=1}^{n} \exp\left(-\frac{\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2}\right) \tag{5}$$

$$= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2}\right) \tag{6}$$

$$= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{\|Y - X\beta\|_2^2}{2\sigma^2}\right) \tag{7}$$

Parameter Estimation

Now, maximising $\log \mathcal{L}(\beta \mid Y, X)$ is equivalent to minimising the sum of the squared error given by:

$$R(\beta) := \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \|Y - X\beta\|_2^2. \qquad (8)$$

Ordinary Least Square (OLS)

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 \qquad (9)$$

Ordinary least squares

The first derivative of the objective function is given by:

$$\frac{\partial R(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\|Y - X\beta\|_2^2\right) \tag{10}$$

$$= \frac{\partial}{\partial \beta}\Big((Y - X\beta)^T(Y - X\beta)\Big) \tag{11}$$

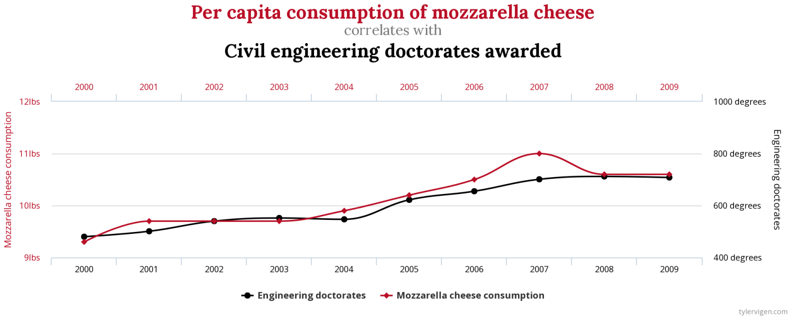
$$= \frac{\partial}{\partial \beta}\Big(Y^TY - 2Y^TX\beta + \beta^TX^TX\beta\Big) \tag{12}$$

$$= -2X^TY + 2X^TX\beta. \tag{13}$$

Therefore equating to zero, we get

- Closed form solution: $\hat{\beta} = (X^TX)^{-1}X^TY$
- But X^TX needs to be invertible

Correlation



Correlation $= 0.95$!

Issues with Correlation

- When the correlation is very high the inverse of $X^T X$ becomes numerically unstable.
- Our linear model may include some variables which are correlated to each other and may lead to overfitting.

To avoid this

- Add some bias in the estimation through penalty term(s)
- Optimise the variance-bias trade-off

Regularisation Methods I

One such regularisation method is penalised regression method.

ℓ_q regularisation [3]

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_q^q \tag{14}$$

where $\|\beta\|_q^q := \sum_{j=1}^{p} |\beta_j|^q$ and $q \leq 1$.

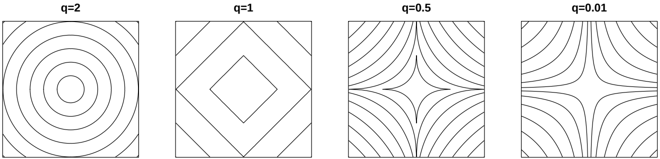


Figure: Regularisation contours for different values of q

Regularisation Methods II

- Gives sparse estimates; that is $\beta_j =$ for some j.
- Analytically NOT solvable and estimation problem becomes non-convex for $q < 1$.

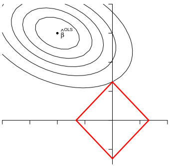


Figure: Relationship between the OLS estimate and the ℓ_1 constraint imposed by the LASSO (red); adapted from [2].

Regularisation Methods III

Ridge Regression [1]

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{15}$$

Similar to OLS method, we can show

$$\frac{\partial}{\partial\beta}\left(R(\beta) + \lambda\|\beta\|_2^2\right) = 2(X^TX + \lambda\mathbf{I}_p)\beta - 2X^TY. \tag{16}$$

Therefore, equating to zero we get,

- $\hat{\beta}(\lambda) = (X^TX + \lambda\mathbf{I}_p)^{-1}X^TY$
- For finding the optimal value of λ, we use cross validation [2].

Prediction

One major aspect of linear regression is to predict new outputs. So, let

$$X^* = \begin{bmatrix} x^*_{11} & \cdots & x^*_{1p} \\ \vdots & & \\ x^*_{r1} & \cdots & x^*_{rp} \end{bmatrix} \tag{17}$$

be r new input then our predicted output is given by:

$$\hat{Y}^* = X^* \hat{\beta}. \tag{18}$$

When we know about the outputs Y^*, we can calculate the prediction error so that

$$Err = \frac{1}{r} \|(Y^* - \hat{Y}^*)\|^2. \tag{19}$$

This is also called as mean squared error.

Bayesian Regression

In previous slides, we learnt about likelihood based approaches for linear regression. Now, we can assume a prior distribution on β to perform Bayesian analyses. We assume that

$$\beta_j \mid \sigma_\beta^2 \sim \mathcal{N}\left(0, \sigma_\beta^2\right) \tag{20}$$

for $j = 1, \cdots, p$ and variance σ_β^2

- This is a natural choice for regression coefficients
- Easy calculations due to conjugacy.

Posterior Calculations

We know from previous classes that

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{21}$$

Then the posterior distribution of β is given by:

$$\begin{aligned}
&P(\beta \mid Y, X) \\
&\propto \mathcal{L}(\beta \mid Y, X) \times P(\beta) && (22) \\
&\propto \exp\left(-\frac{1}{2}(Y - X\beta)^T \Sigma_n^{-1} (Y - X\beta) - \frac{1}{2}\beta^T \Sigma_p^{-1} \beta\right) && (23) \\
&\propto \exp\left(-\frac{1}{2}(\beta - \hat{\beta})^T A (\beta - \hat{\beta})\right) && (24)
\end{aligned}$$

where

- $\Sigma_n = \sigma^2 \mathbf{I}_n$ and $\Sigma_p = \sigma^2 \mathbf{I}_p$
- $\hat{\beta} = A^{-1} X^T \Sigma_n^{-1} y$ and $A = X^T \Sigma_n^{-1} X + \Sigma_p^{-1}$

Posterior Predictive Distribution

Similar to the likelihood based approach, we look into the posterior predictive distribution for the purpose of prediction. Let Y^* be a new point corresponding to new inputs X^* then posterior predictive is given by:

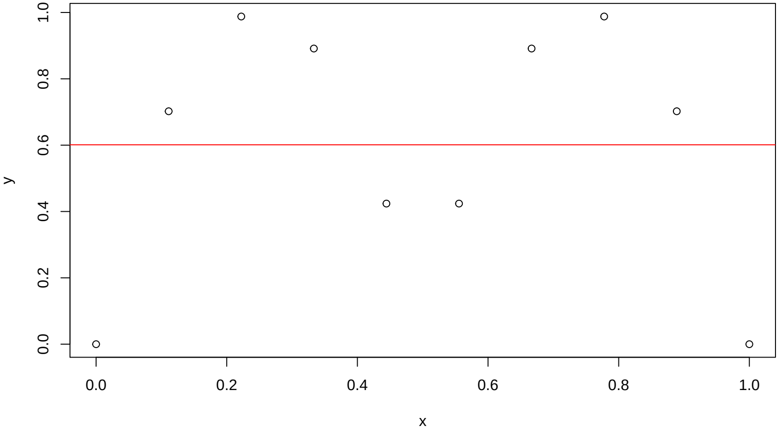
$$P(Y^* \mid Y, X, X^*) = \int_\beta P(Y^* \mid X^*, \beta) P(\beta \mid Y, X) d\beta \tag{25}$$

$$Y^* \mid Y, X, X^* \sim \mathcal{N}\left(X^* \hat{\beta}, X^* A^{-1} {X^*}^T\right) \tag{26}$$

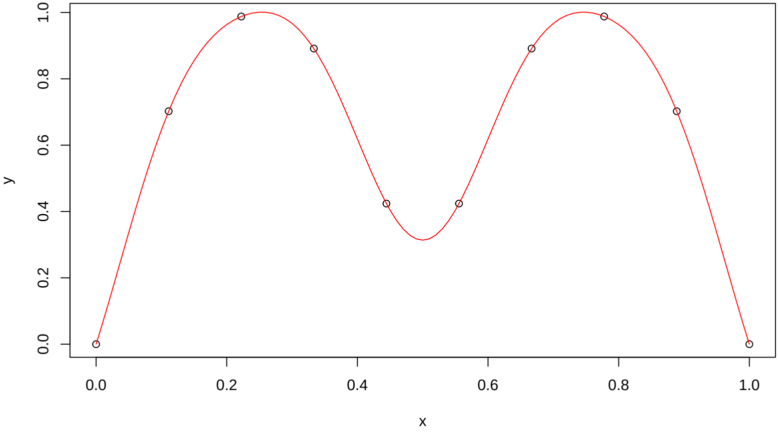
Motivation I

- Classical linear regression models are easy but has limited expressiveness.
- We can project the input parameters at higher dimensions using a set of basis functions.
- We can use these set of basis function and construct a linear model.
  For example, for a variable x, we can have a set of basis $\{1, x, x^2, x^3, \cdots\}$

Motivation II



Motivation III



Gaussian Process

Let x be a p-dimensional input and $\phi(x)$ be the corresponding m-dimensional set of basis functions from $\mathbb{R}^p$ to $\mathbb{R}^m$. Let

$\Phi = (\phi(x_1), \phi(x_2), \cdots, \phi(x_n))^T$ be a $n \times m$ matrix. Then we can define a linear model using these basis functions such that

$$f(x) = \Phi\omega \tag{27}$$

where $\omega = (\omega_1, \cdots, \omega_m)^T$.

Then we can define a Gaussian Process so that,

$$f(x) \sim \mathcal{N}(\Phi\omega, \Sigma_n) \tag{28}$$

Predictive Distribution

The predictive distribution for GP regression is given by:

$$f^* \mid f, X, x^* \sim \mathcal{N}\left(\phi(x^*)A^{-1}\Phi^T\Sigma_n^{-1}f, \phi(x^*)A^{-1}\phi(x^*)^T\right) \quad (29)$$

where $A = \Phi^T\Sigma_n^{-1}\Phi + \Sigma_m^{-1}$.

However, in many cases, $m \gg n$ and inverting A^{-1} can be difficult.

Kernel Trick

Let $K = \Phi\Sigma_m\Phi^T$. Then

$$A \cdot \Sigma_m\Phi^T(K + \Sigma_n)^{-1} \tag{30}$$
$$= (\Phi^T\Sigma_n^{-1}\Phi + \Sigma_m^{-1})\Sigma_m\Phi^T(K + \Sigma_n)^{-1} \tag{31}$$
$$= (\Phi^T\Sigma_n^{-1}\Phi\Sigma_m + \mathbf{I}_m)\Phi^T(K + \Sigma_n)^{-1} \tag{32}$$
$$= \Phi^T(\Sigma_n^{-1}K + \mathbf{I}_n)\Phi^T(K + \Sigma_n)^{-1} \tag{33}$$
$$= \Phi^T(\Sigma_n^{-1}K + \Sigma_n^{-1}\Sigma_n)\Phi^T(K + \Sigma_n)^{-1} \tag{34}$$
$$= \Phi^T\Sigma_n^{-1} \tag{35}$$
$$= A \cdot A^{-1}\Phi^T\Sigma_n^{-1} \tag{36}$$

Therefore, we can replace $A^{-1}\Phi^T\Sigma_n^{-1}$ with $\Sigma_m\Phi^T(K + \Sigma_n)^{-1}$.

Modified Expression

Applying the kernel manipulation we get the following expression

$$f^* \mid f, X, x^* \sim \mathcal{N}\left(\phi(x^*)\Sigma_m \Phi^T (K + \Sigma_n)^{-1} f, V\right) \qquad (37)$$

where $V = \phi(x^*)\Sigma_m \phi(x^*)^T - \phi(x^*)\Sigma_m \Phi^T (K + \Sigma_n)^{-1} \Phi \phi(x^*)^T$.

In Gaussian process can be only defined by the kernel function K and we can set the mean to zero.

Resources

Please find the resource on Gaussian process in the resource section.

References I

[1] A. N. Tikhonov. "On the solution of ill-posed problems and the method of regularization". In: *Dokl. Akad. Nauk SSSR* 151.3 (1963), pp. 501–504. URL: http://www.ams.org/mathscinet-getitem?mr=0162377.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[3] Robert Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.00771.x.