Fundamentals of Machine Learning

LINEAR MODELS

Amit K Roy-Chowdhury

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

Outline

- Logistic Regression
- Linear Regression
- Linear Discriminant Analysis
- Naïve Bayes

Linear Regression

Consider the following equation

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

bias

Slope / weight

If input is 1-D, simple linear regression

$$f(\boldsymbol{x}; \boldsymbol{w}) = ax + b, \qquad b = w_0 \qquad a = w_1$$

If input is N-D, multiple/multivariate linear regression

dimension

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathbf{W}) = \prod_{j=1}^{J} \mathcal{N}(y_j|\boldsymbol{w}_j^\mathsf{T}\boldsymbol{x}, \sigma_j^2) \qquad \boldsymbol{y} \in \mathbb{R}^J$$

Least square regression

To fit a linear regression model to data, we will minimize the negative log likelihood on the training set. The objective function is given by

$$\mathrm{NLL}(\boldsymbol{w}, \sigma^2) = -\sum_{n=1}^{N} \log \underbrace{\left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (y_n - \boldsymbol{w}^\mathsf{T} \boldsymbol{x}_n)^2 \right) \right]}_{\text{Gaussian}} \tag{11.4}$$

$$= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 + \frac{N}{2} \log(2\pi\sigma^2) \tag{11.5}$$

weight

Error

Variance

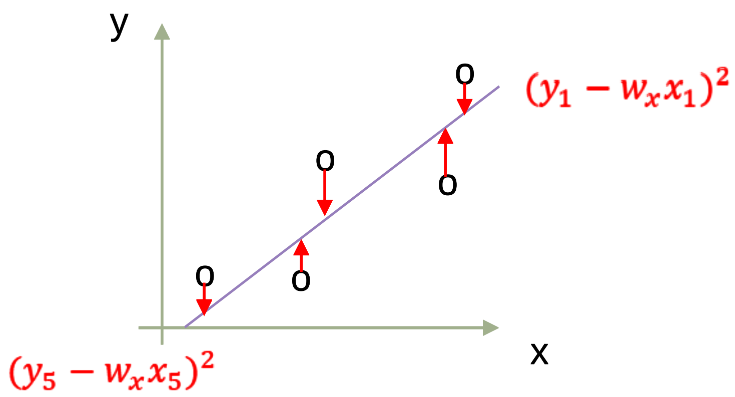
The MLE is the point where $\nabla_{\boldsymbol{w},\sigma} \mathrm{NLL}(\boldsymbol{w}, \sigma^2) = 0$

We can first optimize wrt $\boldsymbol{w}$, and then solve for the optimal σ.

Ordinary least squares – 1D

Residual sum of square is given

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



$(y_1 - w_x x_1)^2$

$(y_5 - w_x x_5)^2$

Least squares solution

Minimize RSS

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

$$\nabla_w \mathrm{RSS}(w) = \mathbf{X}^{\mathsf{T}}\mathbf{X}w - \mathbf{X}^{\mathsf{T}}y \quad \Longrightarrow \quad \hat{w} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y$$

$$\mathbf{H}(w) = \frac{\partial^2}{\partial x^2}\mathrm{RSS}(w) = \mathbf{X}^{\mathsf{T}}\mathbf{X}$$

Hessian is positive semi-definite if X is full rank.

Geometric Interpretation

$$\hat{y} = w_1 x_{:,1} + \cdots + w_D x_{:,D} = \mathbf{X}w$$

X: N rows, D cols, N> D

$$x_{:,d}^{\mathsf{T}}(y - \hat{y}) = 0 \Rightarrow \mathbf{X}^{\mathsf{T}}(y - \mathbf{X}w) = 0 \Rightarrow w = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}y$$

$$\hat{y} = \mathbf{X}w = \boxed{\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}}\,y \longleftarrow \text{Projection Matrix}$$

Geometric Interpretation

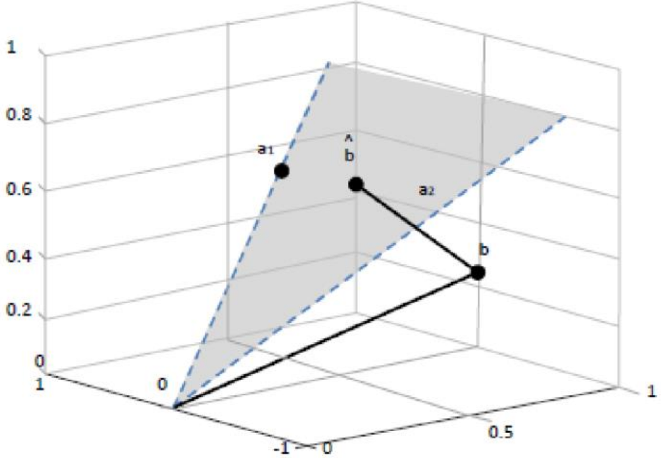


Figure 11.3: Graphical interpretation of least squares for $m = 3$ equations and $n = 2$ unknowns when solving the system $\mathbf{A}x = \mathbf{b}$. a_1 and a_2 are the columns of $\mathbf{A}$, which define a 2d linear subspace embedded in $\mathbb{R}^3$. The target vector $\mathbf{b}$ is a vector in $\mathbb{R}^3$; its orthogonal projection onto the linear subspace is denoted $\hat{\mathbf{b}}$. The line from $\mathbf{b}$ to $\hat{\mathbf{b}}$ is the vector of residual errors, whose norm we want to minimize.

Ridge regression

$$J(w) = (y - \mathbf{X}w)^\mathsf{T}(y - \mathbf{X}w) + \lambda ||w||_2^2$$

$$p(w) = \mathcal{N}(w|\mathbf{0}, \lambda^{-1}\mathbf{I})$$

where $\lambda = \sigma^2/\tau^2$ is the strength of the regularizer

$$\nabla_w J(w) = 2\left(\mathbf{X}^\mathsf{T}\mathbf{X}w - \mathbf{X}^\mathsf{T}y + \lambda w\right)$$

$$\hat{w}_{\text{map}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\mathsf{T}y = (\sum_n x_n x_n^\mathsf{T} + \lambda\mathbf{I}_D)^{-1}(\sum_n y_n x_n)$$

penalizing weights that become too large in magnitude

Lasso Regression

Lasso regression, *RSS* + *λ**|*w*|

Absolute of w

$$||\mathbf{X}w - y||_2^2 + \lambda||w||_1$$

where $||w||_1 \triangleq \sum_{d=1}^{D} |w_d|$

Ridge allows parameters to be small.

Lasso allow parameters to be exactly zero.

This is useful because it can be used to perform feature selection, where the weight of certain features can be zero.

Ridge vs Lasso

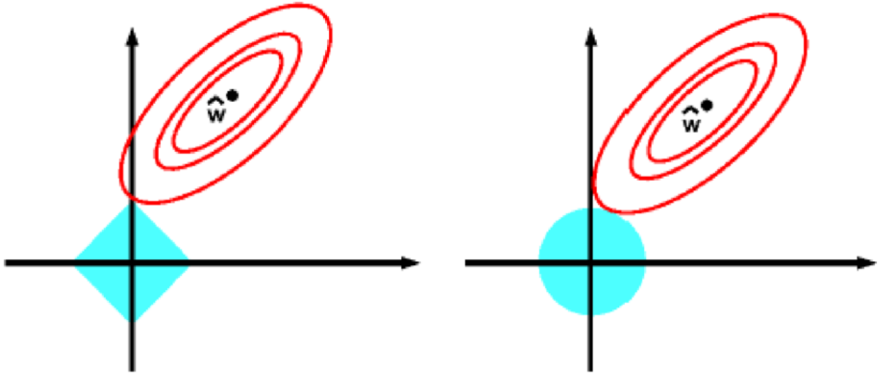


Illustration of ℓ_1 (left) vs ℓ_2 (right) regularization of a least squares problem

Elastic Net

General Equation

$$||\boldsymbol{w}||_q = \left(\sum_{d=1}^{D} |w_d|^q \right)^{1/q}$$

L0 loss

$$||w||_0 = \sum_{d=1}^{D} \mathbb{I}\left(|w_d| > 0\right)$$

L1 loss

$$||\boldsymbol{w}||_1 \triangleq \sum_{d=1}^{D} |w_d|$$

L2 loss

$$||\boldsymbol{w}||_2 \triangleq \sqrt{\sum_{d=1}^{D} |w_d|^2} = \sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$$

Elastic Net – Lasso + Ridge

$$\mathcal{L}(\boldsymbol{w}, \lambda_1, \lambda_2) = ||\boldsymbol{y} - \mathbf{X}\boldsymbol{w}||^2 + \lambda_2 ||\boldsymbol{w}||_2^2 + \lambda_1 ||\boldsymbol{w}||_1$$