# Linear Regression

Jun Du
dr.jundu@gmail.com

Last Updated: January 9, 2018

## 1 Basics

1. Interpretation

   (a) Intuition: Minimizing squared error

   $$E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) \tag{1}$$

   where $\mathbf{X} \in \mathbb{R}^{M \times N}$, $\mathbf{w} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$.

   (b) Probability interpretation:
   Maximum likelihood estimate assuming $p(y_m|\mathbf{x}_m; \mathbf{w}) = \mathcal{N}(y_m|\mathbf{w}^{\mathrm{T}}\mathbf{x}_m, \beta^{-1})$.

   $$\ln \prod_{m=1}^{M} p(y_m|\mathbf{x}_m; \mathbf{w}) = \sum_{m=1}^{M} \ln \mathcal{N}(y_m|\mathbf{w}^{\mathrm{T}}\mathbf{x}_m, \beta^{-1}) \tag{2}$$

   where $y_m \in \mathbb{R}$, $\mathbf{x}_m \in \mathbb{R}^N, \beta \in \mathbb{R}$.

   (c) It can be shown that, minimizing squared error is equivalent to MLE given the Gaussian noise assumption.

2. Analytical solution: normal equations
   Gradient descent can also be applied. When data volume is large, gradient descent is more efficient.

3. Normal equations derivation:

   (a) Setting first derivative to 0

      i. Set the first derivative of $E(\mathbf{w})$ to 0: $\nabla E(\mathbf{w}) = 2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$
      ii. Assume $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible: $\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
      iii. $\mathbf{X}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ have the same null space (see "Introduction to Linear Algebra" P211-212 for proof), hence $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible $\Leftrightarrow$ 0 is the null space of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ $\Leftrightarrow$ 0 is the null space of $\mathbf{X}$ $\Leftrightarrow$ $\mathbf{X}$ has linear independent columns.

(b) Geometric intuition / derivation

$\mathbf{X}\mathbf{w} = \mathbf{y}$ has at least one solution (for $\mathbf{w}$) if and only if $\mathbf{y}$ is in the column space of $\mathbf{X}$. So we formulate $\tilde{\mathbf{y}}$, the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$, such that (1) $\mathbf{X}\hat{\mathbf{w}} = \tilde{\mathbf{y}}$ has at least one solution, and (2) $|\mathbf{y} - \tilde{\mathbf{y}}|$ is minimized.

- Derivation 1:
  i. Using projection matrix (see "Introduction to Linear Algebra" P209-210 for proof), $\tilde{\mathbf{y}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, then we have

$$\begin{aligned}\mathbf{X}\hat{\mathbf{w}} = \tilde{\mathbf{y}} &\Rightarrow \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}\end{aligned}$$

  ii. The rest is the same $\cdots$
- Derivation 2:
  i. Given the projection, $\mathbf{y} - \tilde{\mathbf{y}}$ should be perpendicular to the column space of $\mathbf{X}$, hence we have

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \tilde{\mathbf{y}}) = 0 \Rightarrow \mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = 0 \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

  ii. The rest is the same $\cdots$

(c) Newton-Raphson derivation (see PRML P207 for detail)

  i. The Newton-Raphson update, for minimizing $E(\mathbf{w})$, takes the form

$$\mathbf{w}^{new} = \mathbf{w}^{old} - \mathbf{H}^{-1}\nabla E(\mathbf{w}) \tag{3}$$

  where $\mathbf{H}$ is the Hessian matrix whose elements comprise the second derivatives of $E(\mathbf{w})$ w.r.t. $\mathbf{w}$.

  ii. To minimize squared error, $E(\mathbf{w}) = (\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y})$, then

$$\nabla E(\mathbf{w}) = 2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) \tag{4}$$

$$\mathbf{H} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{5}$$

  iii. The Newton-Raphson update then takes the form

$$\mathbf{w}^{new} = \mathbf{w}^{old} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}^{old} - \mathbf{X}^{\mathrm{T}}\mathbf{y}) \tag{6}$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{7}$$

4. MLE properties

$\mathbf{X}$ is a given / known constant matrix;
$\mathbf{w}$ is an unknown constant vector that is to be estimated (i.e., Frequentest perspective);
$y_m$ is a random variable, where $y_m \sim \mathcal{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_m, \beta^{-1})$;
$\mathbf{y}$ is a random variable vector, where $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I})$;
ML estimator $\hat{\mathbf{w}}$ is a random variable vector, where $\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

(a) $\hat{\mathbf{w}}$ is unbiased.
$\mathrm{E}[\hat{\mathbf{w}}] = \mathrm{E}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{E}[\mathbf{y}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{w}$

(b) $Var(\hat{\mathbf{w}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\beta^{-1}$ (See ESL P47 for proof.)

(c) Given $\mathbf{y}$ has a Gaussian distribution and $\hat{\mathbf{w}}$ is a linear function of $\mathbf{y}$, $\hat{\mathbf{w}}$ also has a Gaussian distribution, $\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\beta^{-1})$.

(d) MLE is also the BLUE (Best Linear Unbiased Estimator). (See ESL P51 for proof.)
That is, among all linear (w.r.t. $\mathbf{y}$) unbiased estimators, MLE has the smallest variance.

5. Loss function interpretation:

   (a) Mean Squared Error (MSE): sensitive to outliers; strict convex

   (b) Mean Absolute Error (MAE): less sensitive to outliers; convex

   (c) $\epsilon$-insensitive Error (Support Vector Regression): less sensitive to outliers

   (d) Huber Loss (Robust Regression; combination of MSE and MAE): less sensitive to outliers.

# 2 Probability Perspective

1. Fundamental assumption: $y_m \sim \mathcal{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_m, \beta^{-1})$

2. Traditional Linear Regression (OLS)

   (a) $\mathbf{w}$ is an unknown constant vector

   (b) $\beta$ is also a constant, but its value is irrelevant to estimating $\mathbf{w}$

   (c) ML for $p(D; \mathbf{w})$ (specifically $p(\mathbf{y}|\mathbf{X}; \mathbf{w})$) is used to optimize $\mathbf{w}$.

   (d) This is equivalent to MSE loss.

3. Ridge and Lasso

   (a) $\mathbf{w}$ is a random variable vector, with prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ for Ridge, and zero mean Laplace prior distribution for Lasso.

   (b) $\alpha$ and $\beta$ are both known constants, or tunable hyperparameters.

   (c) MAP for $p(\mathbf{w}|D)$ is used to optimized $\mathbf{w}$ — still point estimate.

   (d) Ridge tends to shrink the coefficients to small values;
   Lasso tends to shrink the coefficients to zeros (hence can be used for feature selection).

   (e) Ridge is equivalent to MSE loss with $L_2$ regularization;
   Lasso is equivalent to MSE loss with $L_1$ regularization.

4. Bayesian Linear Regression (See PRML Section 3.3 for detail)

    (a) $\mathbf{w}$ is a random variable vector, with prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.
    (b) $\alpha$ and $\beta$ are both known constants, or tunable hyperparameters.
    (c) Full posterior distribution $p(\mathbf{w}|D)$ can be inferred.
    (d) The prediction of $\mathbf{y}$ is:

$$p(\hat{\mathbf{y}}|D, \mathbf{w}; \alpha, \beta) = \int p(\hat{\mathbf{y}}|\mathbf{w}; \beta) p(\mathbf{w}|D; \alpha) d\mathbf{w} \tag{8}$$

    where $\mathbf{X}$ is omitted.

5. Bayesian Linear Regression — Evidence Approximation (See PRML Section 3.5 for detail)

    (a) $\mathbf{w}$ is a random variable vector, with prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.
    (b) $\alpha$ and $\beta$ are unknown, and point estimate is applied to get $\hat{\alpha}$ and $\hat{\beta}$.
        i. $\alpha$ and $\beta$ are regarded unknown constants, and ML for $p(D; \alpha, \beta)$ (by marginalizing $\mathbf{w}$) is used for point estimate.
        ii. $\alpha$ and $\beta$ are regarded random variables with flat prior distributions, and MAP for $p(\alpha, \beta|D)$ is used for point estimate.
        iii. Both end up with the same solution $\hat{\alpha}$ and $\hat{\beta}$.
    (c) Given $\hat{\alpha}$ and $\hat{\beta}$, the full posterior distribution $p(\mathbf{w}|D; \hat{\alpha}, \hat{\beta})$ can be inferred.
    (d) The prediction of $\mathbf{y}$ is:

$$p(\hat{\mathbf{y}}|D, \mathbf{w}; \hat{\alpha}, \hat{\beta}) = \int p(\hat{\mathbf{y}}|\mathbf{w}; \hat{\beta}) p(\mathbf{w}|D; \hat{\alpha}) d\mathbf{w} \tag{9}$$

    where $\mathbf{X}$ is omitted.

6. ARD (Automatic Relevance Determination) for Linear Regression (See MLaPP Section 13.7.1 for detail)

    (a) $\mathbf{w}$ is a random variable vector, with prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$, where $\mathbf{A} = diag(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha} \in \mathbb{R}^N$).
    (b) The prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ is no longer an Isotropic Gaussian (as in Ridge, Lasso, and Bayesian Linear Regression). Each dimension $\mathrm{w}_j$ has its own variance $\alpha_j$.
    (c) $\boldsymbol{\alpha}$ and $\beta$ are random variables with prior distributions: $\alpha_j \sim Ga(a, b)$ and $\beta \sim Ga(c, d)$, where $a$, $b$, $c$, and $d$ are known constants, or tunable hyperparameters.
    (d) MAP for $p(\boldsymbol{\alpha}, \beta)|D)$ is used for point estimate $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$.

(e) Given $\hat{\boldsymbol{\alpha}}$ and $\hat{\beta}$, the full posterior distribution $p(\mathbf{w}|D;\hat{\boldsymbol{\alpha}},\hat{\beta})$ can be inferred.

(f) The prediction of $\mathbf{y}$ is:

$$p(\hat{\mathbf{y}}|D,\mathbf{w};\hat{\boldsymbol{\alpha}},\hat{\beta}) = \int p(\hat{\mathbf{y}}|\mathbf{w};\hat{\beta})p(\mathbf{w}|D;\hat{\boldsymbol{\alpha}})d\mathbf{w} \tag{10}$$

where $\mathbf{X}$ is omitted.

7. Variational Linear Regression — Full Bayesian (See PRML Section 10.3 for detail)

   (a) $\mathbf{w}$ is a random variable vector, with prior distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$.

   (b) $\alpha$ and $\beta$ are also random variables with prior distributions.

   (c) The full posterior distribution $p(\mathbf{w},\alpha,\beta|D)$ can be approximated by VI (Variational Inference), and the marginalized $p(\mathbf{w}|D)$ can also be approximated.

   (d) The prediction of $\mathbf{y}$ is:

$$p(\hat{\mathbf{y}}|D,\mathbf{w},\alpha,\beta) = \int p(\hat{\mathbf{y}}|\mathbf{w},\alpha,\beta)p(\mathbf{w},\alpha,\beta|D)d\alpha d\beta d\mathbf{w} \tag{11}$$

   or equivalently

$$p(\hat{\mathbf{y}}|D,\mathbf{w}) = \int p(\hat{\mathbf{y}}|\mathbf{w})p(\mathbf{w}|D)d\mathbf{w} \tag{12}$$

# 3 Quantile Regression, etc.

1. Quantile Regression

   (a) MSE: $\sum_{m=1}^{M}(y_m - \hat{y}_m)^2$

   (b) MAE: $\sum_{m=1}^{M}|y_m - \hat{y}_m|$

   (c) Quantile regression ($q$ quantile) minimizes a sum that gives *asymmetric* penalties

$$L(\mathbf{w}) = \sum_{m:y_m \geq \mathbf{w}^{\mathrm{T}}\mathbf{x}_m}^{M} q|y_m - \mathbf{w}^{\mathrm{T}}\mathbf{x}_m| + \sum_{m:y_m < \mathbf{w}^{\mathrm{T}}\mathbf{x}_m}^{M} (1-q)|y_m - \mathbf{w}^{\mathrm{T}}\mathbf{x}_m|$$

   that is,
   $q|y_m - \hat{y}_m|$ for *underprediction* (i.e., $y_m \geq \hat{y}_m$)
   $(1-q)|y_m - \hat{y}_m|$ for *overprediction* (i.e., $y_m < \hat{y}_m$)