10-701/15-781, Machine Learning: Homework 1

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on January 21, 2010

1 Linear Regression[30pt, Amr]

In linear regression, we are given training data of the form, $\mathcal{D} = (\mathbf{X}, \mathrm{y}) = \{(\mathrm{x}_i, y_i)\}, i = 1, 2, ..., N$, where $\mathrm{x}_i \in \mathcal{R}^{1\times M}$, i.e. $\mathrm{x}_i = (x_{i,1}, \cdots, x_{i,M})^{\mathrm{T}}$, $y_i \in \mathcal{R}$, $\mathbf{X} \in \mathcal{R}^{N\times M}$, where row i of $\mathbf{X}$ is ${\mathrm{x}_\mathrm{i}}^{\mathrm{T}}$, and $\mathrm{y} = (y_1, \cdots, y_N)^{\mathrm{T}}$. Assuming a parametric model of the form: $y_i = {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta + \epsilon_i$, where ϵ_i are noise terms from a given distribution, linear regression seeks to find the parameter vector β that provides the best of *fit* of the above regression model. One criteria to measure fitness, is to find β that minimizes a given loss function $\mathrm{J}(\beta)$. In class, we have shown that if we take the loss function to be the square-error, i.e.:

$$\mathrm{J}_1(\beta) = \sum_i (y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta)^2 = (\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y})$$

Then

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

Moreover, we have also shown that if we assume that $\epsilon_1, ..., \epsilon_N$ are IID and sampled from the same zero mean Gaussian that is, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then the least square estimate is also the MLE estimate for $p(\mathrm{y}|\mathbf{X}; \beta)$.

In this problem we will explore several extensions to this basic regression model. The following facts might be useful for some parts of problem 2.3:

- The column (row) rank of a matrix A is the maximal number of linearly independent columns (rows) of A.
- If A is $m \times n$ then rank(A)$\leq$ min(n,m)
- An $n \times n$ matrix A is invertible iff it is full-rank, i.e rank(A)=n.
- if A = C^{T}C, then rank(A)=rank(C).
- if A = B + C, then rank(A) $\leq$ rank(B)+ rank(C)

Note: for this problem, you really need to show your work in clear steps to get **full credit**.

1.1 Weighted Least-square

Assume that $\epsilon_1, ..., \epsilon_N$ are independent but each $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

(a) [**2 points**] Write down the formula for calculating the MLE of β.

Solution: $y_i = {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta + \epsilon_i$, thus $p(y_i|\mathrm{x}_i, \beta) = \mathcal{N}({\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta, \sigma_i^2)$. Thus the formula for the MLE of β is:

$$\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta} log \prod_i p(y_i|\mathrm{x}_i, \beta) \\
&= \arg\max_{\beta} \sum_i log\, p(y_i|\mathrm{x}_i, \beta) \\
&= \arg\max_{\beta} \sum_i log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma_i^2})\right) \qquad (2)
\end{aligned}$$

(b) [**4 points**]Calculate the MLE of β. [Show your work]

Solution: Stating form (2)

$$\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta} \sum_i log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma_i^2})\right) \\
&= \arg\max_{\beta} \sum_i log\frac{1}{\sqrt{2\pi\sigma_i^2}} + log\left(\exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma_i^2})\right)
\end{aligned}$$

But first term does not involve β thus we can ignore it.

$$\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta} \sum_i log\left(\exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma_i^2})\right) \\
&= \arg\max_{\beta} \sum_i -\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma_i^2} \\
&= \arg\min_{\beta} \sum_i \frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{\sigma_i^2} \qquad (3)
\end{aligned}$$

Note that we can remove the 2 in the denominator. Now we write the (3) in matrix notation. If we let W be a diagonal matrix with diagonal entry $w_{ii} = \frac{1}{\sigma_i^2}$, we get:

$$\beta_{MLE} = \arg\min_{\beta}(\mathrm{y} - \mathbf{X}\beta)^T\mathbf{W}(\mathrm{y} - \mathbf{X}\beta) \qquad (4)$$

Now we just take derivatives to get β_{MLE} as follows:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\beta}\left((\mathrm{y} - \mathbf{X}\beta)^T\mathbf{W}(\mathrm{y} - \mathbf{X}\beta)\right) \\
&= \frac{\partial}{\partial\beta}\left(\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathrm{y} - \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta - \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathrm{y} + \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta\right)
\end{aligned}$$

For any scalar z, $z = z^T$, therefore, $\left((\beta^T\mathbf{X}^{\mathrm{T}})(\mathbf{W}\mathrm{y})\right)^T = \mathbf{y}^{\mathrm{T}}\mathbf{W}^T\mathbf{X}\beta = \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta$ since $\mathbf{W}^T = \mathbf{W}$ as $\mathbf{W}$ is diagonal. Now putting this back in (5) and taking derivatives, we get:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \beta}\left(\mathbf{y}^{\mathrm{T}}\mathbf{W}\mathrm{y} - 2\beta^{T}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathrm{y} + \beta^{T}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta\right) \\
0 &= -2\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathrm{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta
\end{aligned}
$$

Which means that $\beta_{MLE} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathrm{y})$.

(c) [**2 points**] Show that the MLE you just calculated is the minimizer of the weighted least square loss function $\mathrm{J}_2(\beta) = \sum_i a_i(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2$. Express each a_i in terms of the variance of each example.

Solution: We have already shown that in (3) and form (3) we have $a_i = \frac{1}{\sigma_i^2}$

(d) [**2 points**] Explain why this weighted least-square estimator is preferred to the non-weighted version. (hint: Consider the case when σ_i^2 is large and when it is small).

Solution: When the variance σ_i^2 is high, then the data point (x_i, y_i) might be an outlier as the noise term ϵ_i can be arbitrarily large. In this case, we don't want β_{MLE} to be biased to accommodate such outliers especially when using the squared error loss. The weighted least square formulation in this problem achieves that by weighting the contribution of each data point to the objective function by the inverse of the variance term. Therefore, points with large variance won't contribute much to the loss function and can be safely ignored or at least being given less importance when optimizing for β.

1.2 Laplace noise-model

Assume that $\epsilon_1, ..., \epsilon_N$ are independent and identically distributed according to a Laplace distribution. That is each $\epsilon_i \sim \text{Laplace}(0, b) = \frac{1}{2b}\exp(-\frac{|\epsilon_i|}{b})$.

(a) [**3 points**]Provide the loss function $\mathrm{J}_3(\beta)$ whose minimization is equivalent to finding the MLE of β under the above noise model.

Solution: $y_i = \mathrm{x_i}^{\mathrm{T}}\beta + \epsilon_i$, thus $p(y_i|\mathrm{x}_i, \beta) = \text{Laplace}(\mathrm{x_i}^{\mathrm{T}}\beta, b)$. Thus the formula for the MLE of β is:

$$
\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta} log \prod_i p(y_i|\mathrm{x}_i, \beta) \\
&= \arg\max_{\beta} \sum_i log\, p(y_i|\mathrm{x}_i, \beta) \\
&= \arg\max_{\beta} \sum_i log \left(\frac{1}{2b} \exp(-\frac{|y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|}{b}) \right) \\
&= \arg\max_{\beta} \sum_i log \frac{1}{2b} + log \left(\exp(-\frac{|y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|}{b}) \right) \\
&= \arg\max_{\beta} \sum_i -\frac{|y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|}{b} \\
&= \arg\min_{\beta} \sum_i \frac{|y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|}{b} \\
&= \frac{1}{b} \arg\min_{\beta} \sum_i |y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta| \\
&= \arg\min_{\beta} \sum_i |y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|
\end{aligned}
$$

Thus $J_3(\beta) = \sum_i |y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta|$

(b) [**2 points**] What is the advantage of this model compared to the standard Gaussian assumption? (hint: think about outliers)

Solution: If a point is an outlier then the error in predicting this point given the correct β is much larger in the Gaussian assumption (as it is squared) than in this model (as it is not squared). Therefore, outliers will affect the estimation of β in the Gaussian model more than in the Laplace model. From a modeling point of view, since $y_i = {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta + \epsilon_i$, if y is an outlier, then the model can explain that by making ϵ_i large to accommodate for the difference. This is possible in the Laplace model, since the Laplace distribution has heavier tails than the Gaussian distribution. To relate this to part 1 , to achieve the same effect, we assumed that every example has a different variance. However, these variances have to be estimated (using an EM-like algorithm) since they affect the optimization problem, while in the Laplace model, we don't have to do that. On the other hand, optimizing the L1 loss is harder than optimizing J_2 as the L1 function is not smooth.

1.3 Regularization: Ridge Regression

For this part assume that the noise terms are IID distributed according to $\mathcal{N}(0, \sigma^2)$. Also assume that the number of features M is much larger than the number of training instances N (i.e., $M \gg N$).

(a) [**1 point**] Explain why in this situation, we can NOT compute β according to (1).

Solutoin: In this case, rank$(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ = rank $(\mathbf{X})$ which is smaller than $min(M, N) = N$ which is $<< M$ thus the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which is $M \times M$, is not full rank and thus can not be inverted.

(b) [**5 points**] Instead of minimizing $\mathrm{J}_1(\beta)$, we minimize the following loss function:

$$\mathrm{J}_R(\beta) = \sum_i (y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2 + \lambda \sum_{j=1}^{M} \beta_j^2 = (\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y}) + \lambda \parallel \beta \parallel^2 \tag{5}$$

Derive the value of β^* that minimizes (5) in closed form and show that it is given by $\beta^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{y}$. [please, show you work **in details** to get full credit]

Solution:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\mathrm{J}_R(\beta) &= \frac{\partial}{\partial\beta}\left((\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y}) + \lambda\beta^T\beta\right) \\
&= \frac{\partial}{\partial\beta}\left(\mathbf{y}^{\mathrm{T}}\mathrm{y} - 2\beta^T\mathbf{X}^{\mathrm{T}}\mathrm{y} + \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta + \lambda\beta^T\beta\right) \\
&= -2\mathbf{X}^{\mathrm{T}}\mathrm{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta + 2\lambda\beta
\end{aligned} \tag{6}$$

Equating (6) with 0 and solving for β we get:

$\beta* = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I)^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{y}$. Note that $\lambda\beta = \lambda I\beta$.

(c) [**2 points**] Now revisit your answer to (a) and explain the effect of adding this extra term to the loss function.

Solution: $\mathrm{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I) \leq \mathrm{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \mathrm{rank}(\lambda I) = N + M$. Thus for a proper value of λ, $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I)$ is full rank and can be inverted. To see this, note that if two columns in $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ were linearly dependent, then (λI) adds the same value (λ) but to different components of these columns, thus they become linearly independent in $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I)$.

(d) We have shown in class that the minimizer of $\mathrm{J}_1(\beta)$ is the same as the **MLE** estimate under the IID Gaussian noise assumption. An alternative view is to consider β as a random variable and specify a **prior distribution** $p(\beta)$ on β that expresses our prior belief about the parameters. Then we estimate β using the **MAP (maximum a posteriori)** estimate as:

$$\beta_{\mathrm{MAP}} = \arg\max_{\beta} \prod_{i=1}^{N} p(y_i|\mathrm{x}_i;\beta)\ p(\beta) \tag{7}$$

Assume that $\beta \sim \mathcal{N}(0, \tau^2 I)$:

(i) [**4 points**] Show that maximizing (7) results in the same value of β obtained by minimizing (5) for some value of λ. In other words, show that you can express (7) as (5).

Solution:

$$
\begin{aligned}
\beta_{MAP} &= \arg\max_{\beta} log \prod_i p(y_i|\mathrm{x_i}^{\mathrm{T}}, \beta) p(\beta|\tau) \\
&= \arg\max_{\beta} \sum_i \log p(y_i|\mathrm{x_i}^{\mathrm{T}}, \beta) + \log p(\beta|\tau) \\
&= \arg\max_{\beta} \sum_i log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma^2}) \right) + log \left(\frac{1}{\sqrt{2\pi\tau^2}} \exp\big(-\frac{\beta^T\beta}{2\tau^2}\big) \right) \\
&= \arg\max_{\beta} \sum_i log \frac{1}{\sqrt{2\pi\sigma^2}} + log \left(\exp(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma^2}) \right) + log \frac{1}{\sqrt{2\pi\tau^2}} + \log \left(\exp\big(-\frac{\beta^T\beta}{2\tau^2}\big) \right) \\
&= \arg\max_{\beta} \sum_i \left(-\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma^2} \right) - \frac{\beta^T\beta}{2\tau^2} \\
&= \arg\min_{\beta} \sum_i \left(\frac{(y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2}{2\sigma^2} \right) + \frac{\beta^T\beta}{2\tau^2} \\
&= \frac{1}{2\sigma^2} \arg\min_{\beta} \sum_i (y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2 + \frac{\sigma^2}{\tau^2}\beta^T\beta \\
&= \arg\min_{\beta} \sum_i (y_i - \mathrm{x_i}^{\mathrm{T}}\beta)^2 + \frac{\sigma^2}{\tau^2}\beta^T\beta
\end{aligned}
$$

Which is the same as (5).

(ii) [**1 point**] Express λ as a function of σ and τ.

Solution: $\lambda = \frac{\sigma^2}{\tau^2}$.

(iii) [**2 points**] If we set $\tau = \infty$, what prior belief do we have over β, and how does this affect the loss function?

Solution: Then this is equivalent to a uniform prior belief over all values of β. In other words, all values β are equally likely under this prior. Equivalently, this means that $\lambda = 0$ and the loss function is not regularized at all.