COMP 7070 Advanced Topics in AI and ML January 30, 2024

Lecture 3: Regularization and Logistic Regression

Instructor: Yifan Chen *Scribes: Jin Xiao, Ting Yang* *Proofreader: Zhanke Zhou*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Regularization

Regularization is a technique applied in statistical models to encode preference for simpler solutions and to prevent overfitting by adding a penalty term to the loss function used to fit the model.

3.1.1 Ridge Regression

Ridge Regression, specifically, is a type of linear regression that includes a regularization term. The regularization term added is the $L2$ norm of the coefficients, multiplied by a tuning parameter λ.

$$L = \frac{1}{2n}(Y - Xw)^\top (Y - Xw) + \frac{\lambda}{2}\|w\|^2 \tag{3.1}$$

$$\frac{\partial L}{\partial w} = -\frac{1}{n}X^\top (Y - Xw) + \lambda w = 0 \tag{3.2}$$

the ridge regression estimator is

$$\hat{w} = (X^\top X + n\lambda I)^{-1} X^\top Y \tag{3.3}$$

3.1.1.1 Numeric stability

Observing the equation of $\hat{w}$, we can find that ridge regression addresses numerical stability issues that arise in ordinary least squares (OLS) regression. The condition number is smaller than before.

$$K\left(\left(X^T X + n\lambda I\right)^{-1} X^T\right) \tag{3.4}$$

3.1.1.2 MSE Reduction

Another benefit of ridge regression is the Mean Squared Error (MSE) can be lower than in OLS due to the trade-off between bias and variance. First, we explain what MSE is. In a specific setting of linear regression, for an estimated $\hat{w}$, we can calculate its' MSE easily. In this equation, we use the trace trick.

$$\begin{aligned} MSE(\hat{w}) &= \mathbb{E}(\hat{w} - w^\star)^\top (\hat{w} - w^\star) \\ &= \mathrm{Tr}(\mathbb{E}(\hat{w} - w^\star)(\hat{w} - w^\star)^\top) \\ &= \mathrm{Tr}(M(\hat{w})) \end{aligned} \tag{3.5}$$

In this equation, $M(\hat{w})$ is a matrix of $\hat{w}$; it is a kind of function of estimated $\hat{w}$. We only need to show that the induced matrix $M(\hat{w_\lambda})$ is smaller than $M(\hat{w})$, which means:

$$M(\hat{w_\lambda}) \leq M(\hat{w}) \tag{3.6}$$

The $\leq$ is actually the lower order.

$$M(\hat{w}) = (X^\top X)^{-1} X^\top Y \tag{3.7}$$

We can also induce that :

$$M(\hat{w_\lambda}) \leq M(\hat{w}) \Longrightarrow MSE(\hat{w_\lambda}) \leq MSE(\hat{w}) \tag{3.8}$$

The proof of this is about the PSD matrix:

Firstly, $A \leq B$ means $A - B$ is PSD, which is the definition of lower order. Secondly, for a PSD Matrix, all the diagonals ≥ 0

$$M(\hat{w}_a) - M(\hat{w}) \tag{3.9}$$

$$\text{o.w.} \quad e_i^T(\sim)e_i = \text{i-th} \quad \text{diagonals} \quad < 0. \tag{3.10}$$

This conflicts with the definition of the PSD matrix.

Before we show the complete proof, we can first prove a small proposition:

$$C \cdot A - b \cdot b^T \geq 0 \Leftrightarrow b^T A^{-1} b \leq C \tag{3.11}$$

In which C is a constant, A is an inversible square matrix, and b is a column vector.

To prove it:
Because A is inversible, so to the left hand-side:

$$\begin{aligned} CA^{-\frac{1}{2}}A(A)^{-\frac{1}{2}} - A^{-\frac{1}{2}}bb^\top A^{-\frac{1}{2}} \geq 0 \\ CI - A^{-\frac{1}{2}}bb^\top A^{-\frac{1}{2}} \geq 0 \end{aligned} \tag{3.12}$$

$$x^\top[(CI - A^{-\frac{1}{2}}bb^\top A^{-\frac{1}{2}}]x \geq 0, \quad \forall||x|| = 1 \tag{3.13}$$

$$Cx^T x \geq x^T A^{-\frac{1}{2}} bb^T A^{-\frac{1}{2}} x \tag{3.14}$$

$$\|b^{-1}A^{\frac{1}{2}}\| \leq \sqrt{c}, \quad \text{o.w.} \quad \exists X, \quad x^T \sim X > c \tag{3.15}$$

For a given matrix X of size $m \times n$, SVD decomposes X into:

$$A = U\Sigma V^\top \tag{3.16}$$

$$\|U\Sigma V^\top\| = \sigma\text{max} \Rightarrow V\Sigma^2 V^\top, \quad X = V_1 \tag{3.17}$$

$$\begin{aligned} A &= U\Sigma U^\top \\ A^{-\frac{1}{2}} &= U\Sigma^{-\frac{1}{2}}U^\top \end{aligned} \tag{3.18}$$

3.1.1.3 Bias Vector of Ridge Estimator

The bias of an estimator is the difference between its expected value and the true parameter value. For the ridge estimator, the bias can be calculated as:

$$\begin{aligned} M(\hat{w}_\lambda) &= \text{Var}\left(\hat{w}_\lambda\right) + \text{E}\left(\hat{w}_\lambda - \hat{w}^\star\right) \cdot \text{E}\left(\hat{w}_\lambda - \hat{w}^\star\right)^\top \\ \text{E}\left(X^\top X + n\lambda I\right)^{-1} X^\top Y (= Xw^\star + E) & \\ &= (\text{X}^\top X + n\lambda I)^{-1} X^\top X w^\star \\ &= (\text{X}^\top X + n\lambda I)^{-1}\left(X^\top X + n\lambda I - n\lambda I\right) w^\star \\ &= w^\star - n\lambda\left(X^\top X + n\lambda I\right)^{-1} w^\star \end{aligned} \tag{3.19}$$

So,

$$Bias = n\lambda\left(X^\top X + n\lambda I\right)^{-1} w^\star \tag{3.20}$$

Notice that $(X^\top X + \lambda I)^{-1} X^\top X$ is not the identity matrix, resulting in a non-zero bias when $\lambda > 0$.

3.1.1.4 Variance of Ridge Estimator

First we define a matrix B_λ as:

$$B_\lambda = (X^\top X + n\lambda I)^{-1} X^\top X \tag{3.21}$$

Then

$$\hat{w}_\lambda = B_\lambda \left(X^\top X\right)^{-1} X^\top Y = B\lambda\hat{w} \tag{3.22}$$

$$\begin{aligned} \mathrm{Var}\left(\hat{w}_\lambda\right) &= B_\lambda \mathrm{Var}(\hat{w}) B_\lambda^\top \\ &= \sigma^2 B_\lambda (X^\top X)^{-1} B_\lambda^\top \\ &= \sigma^2 (X^\top X + naI)^{-1} X^\top X (X^\top X + n\lambda I)^{-1} \end{aligned} \tag{3.23}$$

Next we need to compare the $\mathrm{Var}(\hat{w})$ and $\mathrm{Var}(\hat{w}_a)$
A small trick to simplify the decompose to matrix: rewrite the $X^\top X^{-1}$ as $U\Sigma^{-1}U^\top$

$$\begin{aligned} &\mathrm{Var}(\hat{w}) - \mathrm{Var}(\hat{w}_a) \\ &= \sigma^2 U\Sigma^{-1}U^\top - \sigma^2 U[(\Sigma + n\lambda I)\Sigma^{-1}(\Sigma + n\lambda I)]U^\top \\ &= \sigma^2 U[\Sigma + n\lambda I)^{-1}(\Sigma + n\lambda I)\Sigma^{-1}(\Sigma + n\lambda I(\Sigma + n\lambda I)^{-1}]U^T \\ &= \sigma^2 U[\Sigma + n\lambda I)^{-1}(2n\lambda I + n^2\lambda^2\Sigma^{-1})(\Sigma + n\lambda I)^{-1}]U^T \\ &= \sigma^2 (X^\top X + n\lambda I)^{-1}(2n\lambda I + n^2\lambda^2 (X^\top X)^{-1})((X^\top X + n\lambda I)^{-1})] \end{aligned} \tag{3.24}$$

To show the MSE of the ridge estimator $\hat{w}_\lambda$ is better, we need to return to the X form and insert the bias into the difference. So the $\mathrm{Var}(\hat{w}) - Var(\hat{w}_a)$ can be write as

$$\sigma^2 (X^\top X + n\lambda I)^{-1}(2n\lambda I + n^2\lambda^2 (X^\top X)^{-1})((X^\top X + n\lambda I)^{-1})] - bias_\lambda bias_\lambda^\top \tag{3.25}$$

in which,

$$bias_\lambda bias_\lambda^\top = n^2\lambda^2 (X^T X + n\lambda I)^{-1} w^* w^{*\top} (X^\top X + n\lambda I)^{-1} \tag{3.26}$$

$$\mathrm{Var}(\hat{w}) - \mathrm{Var}(\hat{w}_a) = \sigma^2 \cdot 2n\lambda I + n^2\lambda^2\sigma^2 (X^\top X)^{-1} - n^2 a^2 w^* w^{*\top} \tag{3.27}$$

It is sufficient to show

$$2\sigma^2 I - n\lambda w^* w^{*\top} \geq 0 \tag{3.28}$$

We can finally get the following:

$$w^{*\top} w^* \leq \frac{2t^2}{\eta\lambda} \quad \Longleftrightarrow \lambda \leq \frac{1}{n} 2\sigma^2 \cdot \frac{1}{||w^\star||^2} \tag{3.29}$$

Overall, the MSE, which decomposes into the sum of the variance and the square of the bias, may decrease if the increase in bias is offset by a larger decrease in variance.

3.2 Logistic Regression

3.2.1 Differentiate Terms

3.2.1.1 Softmax v.s. Softmax for X_{1-n}

The softmaximum of $\vec{X}$ is defined by:

$$\ln \sum_{i=1}^{n} e^{x_i} \approx \ln e^{x_{max}} = x_{max} \tag{3.30}$$

The softmax function σ takes as input a vector $\vec{X}$ of n real numbers and normalizes it into a probability distribution consisting of n probabilities proportional to the exponentials of the input numbers. It approximates a one-hot vector $\vec{y} = (0, 0, \ldots, 1, \ldots, 0, 0)$, is defined by the formula:

$$\begin{aligned} \sigma(\vec{X}) &= \left\{ \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} \right\}_{i=1}^{n} \\ &= \exp(\vec{X}) / \langle \exp(\vec{X}), I_n \rangle \end{aligned} \tag{3.31}$$

3.2.2 Cross Entropy

The cross-entropy $H(p,q)$ represents the expected value of the information content $I_p(x)$ under the distribution q. Here, p and q are two probability distributions, with q representing the ground truth. The formula is defined as follows:

$$\begin{aligned} H(p,q) &= \mathbb{E}_q\left[I_p(x)\right] \\ &= \mathbb{E}_q\left[-\log p(x)\right] \\ &= \langle y, -\log \sigma(x)\rangle \end{aligned} \tag{3.32}$$

Cross entropy and negative log-likelihood are equivalent, which can be represented as:

$$\log \prod_{i=1}^{n} \left[\sigma(\vec{X})\right]_i^{y_i} \tag{3.33}$$

In PyTorch, cross-entropy loss implicitly combines cross-entropy computation with softmax. The input provided is the unnormalized logits.

3.2.3 Logits

The logit function is defined as $\log\left(\frac{p}{1-p}\right)$. In the special case of two classes, p can be calculated as follows:

$$p = \sigma\left(X_1\right) = \frac{\exp\left(X_1\right)}{1+\exp\left(X_1\right)} \tag{3.34}$$

where σ is the sigmoid function. Then

$$\log \frac{p}{1-p} = \log \exp\left(X_1\right) = X_1 \tag{3.35}$$

3.2.4 Other Form of Logistic Regression

In some cases, the logistic regression is represented as follows:

$$\frac{\exp\left(W_i^\top X\right)}{1+\sum_{j=1}^{n-1}\exp\left(W_j^\top X\right)} \tag{3.36}$$

Since $W_n = 0$, the cross-entropy form is overparameterized, and the Hessian matrix is singular.

3.2.5 Numerical Issue

The softmax function may exhibit overflow and underflow numerical issues. For instance, the expression $\sum \exp(x_i)$ can lead to overflow when $x_1 = 1000$, $x_2 = 2000$, and $x_3 = 3000$. Additionally, the modified expression $\frac{\exp(x_i - x_{max})}{\sum \exp(x_j - x_{max})}$ may lead to underflow. Log softmax is advantageous over softmax for improved numerical performance and gradient optimization.

3.2.6 Other Form of Cross-Entropy Loss

In binary classification, the ground truth can be $y \in \{0,1\}$ or $y \in \{-1,1\}$. The cross-entropy loss is defined as follows:

$$e\left(z\right) = \log\left(1+\exp\left(-yz\right)\right) \tag{3.37}$$

Since

$$\sigma\left(z\right) = \frac{\exp\left(z\right)}{1+\exp\left(z\right)} \tag{3.38}$$

where σ is the sigmoid function. For instance, when $P\left(Y=1|z\right)$, we can get $\sigma\left(z\right) = \frac{1}{1+\exp(-z)}$. When $P\left(Y=-1|z\right)$, we can get $\sigma\left(z\right) = \frac{1}{1+\exp(z)}$. Since

$$P\left(Y=y|z\right) = \sigma\left(yz\right) \tag{3.39}$$

e is the negative log-likelihood, which can be obtained:

$$e\left(z\right) = -\ln \sigma\left(yz\right) = \ln\left(1+\exp\left(-yz\right)\right) \tag{3.40}$$

3.2.7 Convexity

In this section, we will explain why logistic regression is a convex function. According to the conclusion from the second lecture about logistic regression:

$$\frac{\partial l}{\partial a} = -y + \text{softmax}\left(a\right), a = WX \tag{3.41}$$

Then represent the Hessian matrix $H = \frac{\partial^2 l}{\partial^2 a}$ column by column:

$$H_i = \frac{\partial g_i}{\partial a} = \frac{\partial \langle e_i, \frac{\partial l}{\partial a} \rangle}{\partial a} \tag{3.42}$$

Then

$$\begin{aligned} \partial g_i &= \partial \exp\left(\langle e_i, \text{logsoftmax}\left(a\right)\rangle\right) \\ &= \langle e_i, \text{softmax}\left(a\right)\rangle \left[e_i - \text{softmax}\left(a\right)\right]^\top \partial a \end{aligned} \tag{3.43}$$

where $\langle e_i, \text{softmax}\left(a\right)\rangle$ is denoted as σ_i. Therefore

$$H_i = \frac{\partial g_i}{\partial a} = \sigma_i \left(e_i - \sigma\right) \tag{3.44}$$

The Hessian matrix can be represented as follows:

$$H = \text{diag}\left(\sigma\right) - \sigma\sigma^\top \tag{3.45}$$

For $\forall x$ and x is a nonzero vector. We can get

$$x^\top H x = \sum \sigma_i x_i^2 - \left(\sigma_i x_i\right)^2 >= 0 \tag{3.46}$$

where $\sum \sigma_i = 1$. Therefore, the Hessian matrix is a positive semidefinite matrix, and the logistic regression is a convex problem.