HW 1

Karl Hiner

September 4, 2023

1 Linear Regression

1.1 a

$$\hat{w} = (X^T X)^{-1} X^T Y \tag{1}$$

$$y_i = w^T x_i + \varepsilon_i, \tag{2}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $w \in \mathbb{R}^d$, and $\{X^i, Y^i\}$ is the i-th data point, with $1 \le i \le m$.

Using the normal equation (Eqn. 1), and the model (Eqn. 2), derive the expectation $\mathbb{E}[\hat{w}]$. Note that here X is fixed, and only Y is random.

$$\begin{aligned}
\mathbb{E}[\hat{w}] &= \mathbb{E}\big[(X^T X)^{-1} X^T Y\big] && \text{Eqn. 1} \\
&= \mathbb{E}\big[(X^T X)^{-1} X^T (Xw + \varepsilon)\big] && \text{Substitute } Y \\
&= \mathbb{E}\big[(X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Distribute} \\
&= \mathbb{E}\big[w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Simplify} \\
&= w + (X^T X)^{-1} X^T \mathbb{E}[\varepsilon] && \text{Linearity of expectation} \\
&= w && \text{Since } \mathbb{E}[\varepsilon] = 0
\end{aligned}$$

1.2 b

Similarly, derive the variance $\mathrm{Var}[\hat{w}]$.

$$\begin{aligned}
\mathrm{Var}[\hat{w}] &= \mathrm{Var}\big[(X^T X)^{-1} X^T (Xw + \varepsilon)\big] && \text{Eqn. 1, Substitute } Y \\
&= \mathrm{Var}\big[(X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Distribute} \\
&= \mathrm{Var}\big[w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Simplify} \\
&= \mathrm{Var}\big[(X^T X)^{-1} X^T \varepsilon\big] && \text{Since } w \text{ is constant} \\
&= (X^T X)^{-1} X^T \mathrm{Var}[\varepsilon] X (X^T X)^{-1} && \mathrm{Var}[\boldsymbol{b}^T \boldsymbol{X}] = \boldsymbol{b}^T \mathrm{Var}[\boldsymbol{X}] \boldsymbol{b} \\
&= (X^T X)^{-1} X^T (\sigma^2 I_m) X (X^T X)^{-1} && \mathrm{Var}[\varepsilon] \triangleq \sigma^2 \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} && \text{Commute } \sigma^2 \\
&= \sigma^2 (X^T X)^{-1} && \text{Simplify}
\end{aligned}$$

1.3 c

Under the white noise assumption above, does $\hat{w}$ follow a Gaussian distribution with mean and variance in (a) and (b), respectively? Why or why not?

Answer: Yes, $\hat{w}$ follows a Gaussian distribution with mean and variance in (a) and (b), respectively. This is because $\hat{w}$ is a linear combination of the random variables ε_i, which are Gaussian by assumption. Since $\hat{w}$ is a linear combination of Gaussian random variables, it is itself Gaussian, with $\hat{w} \sim \mathcal{N}(w, \sigma^2(X^TX)^{-1})$.

1.4 d: Weighted linear regression

Suppose we keep the independence assumption but remove the same variance assumption. In other words, data points would be still sampled independently, but now they may have different variance σ_i. Thus, the variance (the covariance matrix) of ε would still be diagonal, but with different values:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{bmatrix}$$

Derive the estimator $\hat{w}$ (similar to the normal equations) for this problem using matrix-vector notations with Σ.

Answer:

We want to minimize

$$\arg\min_w \sum_{i=1}^{m} \frac{1}{\sigma_i^2}(y_i - w^Tx_i)^2.$$

In matrix-vector notation, this is equivalent to

$$\arg\min_w (Y - Xw)^T\Sigma^{-1}(Y - Xw).$$

Taking the derivative with respect to w and setting it to zero:

$$\begin{aligned} \frac{\partial}{\partial w}\big((Y - Xw)^T\Sigma^{-1}(Y - Xw)\big) &= 0 \\ \frac{\partial}{\partial w}\big(w^TX^T\Sigma^{-1}Xw - 2w^TX^T\Sigma^{-1}Y + Y^T\Sigma^{-1}Y\big) &= 0 \\ -2X^T\Sigma^{-1}Y + 2X^T\Sigma^{-1}Xw &= 0 \\ (X^T\Sigma^{-1}X)w &= X^T\Sigma^{-1}Y \end{aligned}$$

Thus, the weighted least squares estimator $\hat{w}$ is:

$$\hat{w} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

2 Ridge Regression

For linear regression, it is often assumed that $y_i = w^Tx_i + \varepsilon$, where $w, x \in \mathbb{R}^d$ by absorbing the constant term (bias) in an affine hypothesis into w, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. Given m i.i.d. samples $z_i = (x_i, y_i), 1 \leq i \leq m$, we define $Y = (y_1, \cdots, y_m)^T$ and

$X = (x_1, \cdots, x_m)^T$. Thus, we have $Y \sim \mathcal{N}(Xw, \sigma^2 I_m)$. Show that the ridge regression estimate is the mean of the posterior distribution under a Gaussian prior $w \sim \mathcal{N}(0, \tau^2 I)$. Find the explicit relation between the regularization parameter λ in the ridge regression estimate of the parameter w, and the variances σ^2, τ^2.

Answer:

The ridge regression estimate is defined as

$$\begin{aligned}\hat{w}^{\text{Ridge}} &\triangleq \arg\min_w \sum_{i=1}^m (w^T x_i - y_i)^2 + \lambda \|w\|^2 \\ &= \arg\min_w \|Xw - Y\|^2 + \lambda \|w\|^2.\end{aligned}$$

Taking the derivative with respect to w and setting it to zero results in the following expression for $\hat{w}^{\text{Ridge}}$ (as derived in class and in the text):

$$\hat{w}^{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T Y$$

Now, we'll show that this estimator is also the mean of the posterior distribution of w when we assume a Gaussian prior $w \sim \mathcal{N}(0, \tau^2 I)$.

The posterior distribution of w is proportional to the product of the likelihood and the prior:

$$\begin{aligned}p(w|Y) &\propto p(Y|w)p(w) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|Y - Xw\|^2\right)\exp\left(-\frac{1}{2\tau^2}\|w - 0\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|Xw - Y\|^2\right)\exp\left(-\frac{1}{2\tau^2}\|w\|^2\right)\end{aligned}$$

(We neglect the normalization constant $P(Y)$ since it does not depend on w. We also neglect both of the Gaussian normalizing factors since they do not affect the location of the mode.)

We want to find the mean of this posterior distribution. Since multiplying two Gaussian PDFs results in another Gaussian PDF, and since the mode of a Gaussian PDF is also its mean, we can find the mean of this posterior by taking its negative and settings its derivative with respect to w to zero. This value, $\hat{w}^{\text{Mean}}$, will be the (single) maximum of the (Gaussian) posterior distribution:

$$\begin{aligned}-log(p(w|Y)) &= -log\left(\exp\left(-\frac{1}{2\sigma^2}\|Xw - Y\|^2\right)\exp\left(-\frac{1}{2\tau^2}\|w\|^2\right)\right) \\ &= \frac{1}{2\sigma^2}\|Xw - Y\|^2 + \frac{1}{2\tau^2}\|w\|^2 \\ 0 &= \frac{\partial}{\partial w}\left(\frac{1}{2\sigma^2}\|Xw - Y\|^2 + \frac{1}{2\tau^2}\|w\|^2\right) \\ 0 &= \frac{1}{\sigma^2}X^T(Xw - Y) + \frac{1}{\tau^2}w \\ 0 &= \frac{1}{\sigma^2}X^T Xw - \frac{1}{\sigma^2}X^T Y + \frac{1}{\tau^2}w \\ \hat{w}^{\text{Mean}} &= \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^T X + \frac{1}{\tau^2}I\right)^{-1} X^T Y\end{aligned}$$

Now, we can find the value of λ that makes $\hat{w}^{\text{Mean}}$ equal to $\hat{w}^{\text{Ridge}}$:

$$\begin{aligned}
\hat{w}^{\text{Mean}} &= \hat{w}^{\text{Ridge}} \\
\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I\right)^{-1} X^TY &= (X^TX + \lambda I)^{-1}X^TY \\
\left(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I\right)^{-1} \frac{1}{\sigma^2} &= (X^TX + \lambda I)^{-1} \\
\frac{1}{\sigma^2} &= \left(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I\right)(X^TX + \lambda I)^{-1} \\
\frac{1}{\sigma^2}(X^TX + \lambda I) &= \left(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I\right) \\
\frac{1}{\sigma^2}X^TX + \frac{\lambda}{\sigma^2}I &= \frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I \\
\frac{\lambda}{\sigma^2}I &= \frac{1}{\tau^2}I \\
\lambda &= \frac{\sigma^2}{\tau^2}
\end{aligned}$$

Thus, $\hat{w}^{\text{Mean}} = \hat{w}^{\text{Ridge}}$ if we set $\lambda = \frac{\sigma^2}{\tau^2}$.

3 Lasso estimator

The LASSO regression problem can be shown to be the following optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{m} (\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 \text{ subject to } \|\boldsymbol{w}\|_1 \leq \lambda,$$

where $\lambda > 0$ is a regularization parameter. Here, we develop a stochastic gradient descent (SDG) algorithm for this problem, which is useful when we have $m >> d$, where d is the dimension of the parameter space.

3.1 a

Write $\boldsymbol{w} = \boldsymbol{w}^+ - \boldsymbol{w}^-$, where $\boldsymbol{w}^+, \boldsymbol{w}^- \geq 0$ are the positive and negative parts of $\boldsymbol{w}$ respectively. Let w_j be the jth component of $\boldsymbol{w}$. When $w_j \leq 0$, $w_j^+ = 0$ and $w_j^- = -w_j$. Similarly, when $w_j \geq 0$, $w_j^+ = w_j$ and $w_j^- = 0$. Find a quadratic function, Q, of $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ such that

$$\min_{\boldsymbol{w}^+,\boldsymbol{w}^-\geq 0} \lambda \sum_{i=1}^{m} Q(\boldsymbol{w}^+, \boldsymbol{w}^-)$$

is equivalent to the above LASSO problem. Expain the equivalence.

Answer:

[Mohri et al] show in Eqn. 11.33 that the LASSO problem can be rewritten as

$$\min_{\boldsymbol{w}^+,\boldsymbol{w}^-\geq 0} \sum_{i=1}^{m} \big((\boldsymbol{w}^+ - \boldsymbol{w}^-)\boldsymbol{x}_i - y_i\big)^2 + \lambda \sum_{j=1}^{d} (w_j^+ + w_j^-).$$

We can rewrite this in the required quadratic form as follows:

$$\min_{\boldsymbol{w}^+,\boldsymbol{w}^-\geq 0} \lambda\left(\sum_{i=1}^{m}\frac{1}{\lambda}\big((\boldsymbol{w}^+-\boldsymbol{w}^-)\boldsymbol{x}_i-y_i\big)^2+\sum_{j=1}^{d}(w_j^+ + w_j^-)\right)$$

$$\min_{\boldsymbol{w}^+,\boldsymbol{w}^-\geq 0} \lambda\sum_{i=1}^{m}\left(\frac{1}{\lambda}\big((\boldsymbol{w}^+-\boldsymbol{w}^-)\boldsymbol{x}_i-y_i\big)^2+\frac{1}{m}\sum_{j=1}^{d}(w_j^+ + w_j^-)\right)$$

In this form, we can see that

$$Q(\boldsymbol{w}^+,\boldsymbol{w}^-)=\frac{1}{\lambda}\big((\boldsymbol{w}^+-\boldsymbol{w}^-)\boldsymbol{x}_i-y_i\big)^2+\frac{1}{m}\sum_{j=1}^{d}(w_j^+ + w_j^-).$$

3.2 b

[Mohri et al Ex. 11.10] Derive a stochastic gradient descent algorithm for the quadratic program (with affine constraints) in part (a).

3.3 c

Suppose $X=[x_1,\cdots,x_m]^T$ is orthonormal and there exists a solution w for $Xw=Y$, where $Y=[y_1,\cdots,y_m]^T$ with no more than k non-zero elements. Can the SGD algorithm get arbitrarily close to w? Explain why or why not.

4 Logistic Regression

Logistic regression is named after the log-odds of success (the logit of the probability) defined as below:

$$\ln\left(\frac{P[Y=1|X=x]}{P[Y=0|X=x]}\right),$$

where

$$P[Y=1|X=x]=\frac{\exp(w_0+w^Tx)}{1+\exp(w_0+w^Tx)}.$$

4.1 a

Show that log-odds of success is a linear function of X.

4.2 b

Show that the logistic loss $L(z)=\log(1+\exp(-z))$ is a convex function.

5 Programming: Recommendation System

Problem Summary: A rating by user u on movie i is approximated by

$$M_{u,i}\approx\sum_{k=1}^{d}U_{u,k}V_{i,k}. \tag{3}$$

We want to fit each element of U and V by minimizing squared reconstruction error over all training data points. That is, the objective function we minimize is given by

$$E(U,V)=\sum_{u=1}^{n}\sum_{i=1}^{m}(M_{u,i}-U_uV_v^T)^2=\sum_{u=1}^{n}\sum_{i=1}^{m}(M_{u,i}-\sum_{k=1}^{d}U_{u,k}V_{i,k})^2, \tag{4}$$

where U_u is the uth row of U and V_i is the ith row of V.

We use gradient descent:

$$U_{v,k} \leftarrow U_{v,k} - \mu \frac{\partial E}{\partial U_{v,k}}, \quad V_{j,k} \leftarrow V_{j,k} - \mu \frac{\partial E}{\partial V_{j,k}}, \tag{5}$$

where μ is the update rate.

5.1 a

Derive the update formula in (5) by solving the partial derivatives.

Answer:

$$\begin{aligned}
\frac{\partial E}{\partial U_{v,k}} &= \frac{\partial}{\partial U_{v,k}} \sum_{u=1}^{n} \sum_{i=1}^{m} (M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'})^2 \\
&= \sum_{u=1}^{n} \sum_{i=1}^{m} \frac{\partial}{\partial U_{v,k}} (M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'})^2 \\
&= \sum_{u=1}^{n} \sum_{i=1}^{m} 2(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'}) \frac{\partial}{\partial U_{v,k}} (M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'}) && \text{Chain rule} \\
&= \sum_{u=1}^{n} \sum_{i=1}^{m} 2(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'})(-\frac{\partial}{\partial U_{v,k}} \sum_{k'=1}^{d} U_{u,k'} V_{i,k'}) && \tfrac{\partial}{\partial U_{v,k}} M_{u,i} = 0 \\
&= \sum_{i=1}^{m} 2(M_{v,i} - \sum_{k'=1}^{d} U_{v,k'} V_{i,k'})(-V_{i,k}) && \tfrac{\partial}{\partial U_{v,k}} U_{u,k'} V_{i,k'} = 0, k' \neq k, u \neq v \\
&= -2 \sum_{i=1}^{m} (M_{v,i} - \sum_{k'=1}^{d} U_{v,k'} V_{i,k'}) V_{i,k}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial E}{\partial V_{j,k}} &= \frac{\partial}{\partial V_{j,k}} \sum_{u=1}^{n} \sum_{i=1}^{m} (M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'})^2 \\
&= \sum_{u=1}^{n} \sum_{i=1}^{m} 2(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'} V_{i,k'})(-\frac{\partial}{\partial V_{j,k}} \sum_{k'=1}^{d} U_{u,k'} V_{i,k'}) && \text{Same first three steps} \\
&= \sum_{u=1}^{n} 2(M_{u,j} - \sum_{k'=1}^{d} U_{u,k'} V_{j,k'})(-U_{j,k}) && \tfrac{\partial}{\partial V_{j,k}} U_{u,k'} V_{j,k'} = 0, k' \neq k, i \neq j \\
&= -2 \sum_{u=1}^{n} (M_{u,j} - \sum_{k'=1}^{d} U_{u,k'} V_{j,k'}) U_{j,k}
\end{aligned}$$

5.2 b

Redo part (a) using the regularized objective function below.

$$E(U,V) = \sum_{u=1}^{n} \sum_{i=1}^{m} (M_{u,i} - \sum_{k=1}^{d} U_{u,k} V_{i,k})^2 + \mu \sum_{u,k} U_{u,k}^2 + \lambda \sum_{i,k}^{m} V_{i,k}^2$$

$$\begin{aligned}\frac{\partial E}{\partial U_{v,k}} &= \frac{\partial}{\partial U_{v,k}}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'}V_{i,k'})^2 + \mu\sum_{u=1}^{n}\sum_{k'=1}^{d} U_{u,k'}^2 + \lambda\sum_{i=1}^{m}\sum_{k'=1}^{d} V_{i,k'}^2\right) \\ &= \frac{\partial}{\partial U_{v,k}}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'}V_{i,k'})^2\right) + \frac{\partial}{\partial U_{v,k}}\mu\sum_{u=1}^{n}\sum_{k'=1}^{d} U_{u,k'}^2 + \frac{\partial}{\partial U_{v,k}}\lambda\sum_{i,k}^{m} V_{i,k}^2 \\ &= -2\sum_{i=1}^{m}(M_{v,i} - \sum_{k'=1}^{d} U_{v,k'}V_{i,k'})V_{i,k} + 2\mu U_{v,k}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial V_{j,k}} &= \frac{\partial}{\partial V_{j,k}}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}(M_{u,i} - \sum_{k'=1}^{d} U_{u,k'}V_{i,k'})^2\right) + \frac{\partial}{\partial U_{v,k}}\mu\sum_{u,k} U_{u,k}^2 + \frac{\partial}{\partial U_{v,k}}\lambda\sum_{i=1}^{m}\sum_{k'=1}^{d} V_{i,k'}^2 \\ &= -2\sum_{u=1}^{n}(M_{u,j} - \sum_{k'=1}^{d} U_{u,k'}V_{j,k'})U_{j,k} + 2\lambda V_{j,k}\end{aligned}$$