Solutions to Homework 3

Nguyen Trung Hai

Sep 17, 2019

Problem 1

(a)

Experience E is the set of training data which includes about 55M rows of data. Each row corresponds to a taxi ride that has input features (`pickup_datetime`, `pickup_longitude`, `pickup_latitude`, `dropoff_longitude`, `dropoff_latitude`, `passenger_count`) and target `fare_amount`. The class of the tasks T of the algorithms used to solve this problem is prediction of a numerical target value given the input features.

(b)

We can use either RMSE (Root Mean Square Error) or R_2 (R-squared) calculated on the set of test data as a performance measure P.

RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{1}$$

where y_i and $\hat{y}_i$ are the true and predicted values, respectively, of the target variable of test data point i. RMSE is the standard deviation of the unexplained residuals $\epsilon_i = y_i - \hat{y}_i$. The lower RMSE the better a model is. An advantage of RMSE is that it has the same unit as y, so RMSE gives an absolute measure of how good a model is in predicting the target.

R-squared is defined as

$$R_2 = \frac{\text{TSS}}{\text{TSS} - \text{RSS}}, \tag{2}$$

where the *total sum of squares* (TSS) is given by

$$\text{TSS} = \sum_{i=1}^{N}(y_i - \bar{y})^2, \tag{3}$$

where $\bar{y}$ is the mean value of y. The *residual sum of squares* (RSS) is given by

$$\text{RSS} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{4}$$

R_2 is interpreted as the fraction of variance explained by the model. It ranges from zero to one, so it gives a relative measure of how good a model is. Zero R_2 means that the model being considered does not improve the prediction over the mean model (the model that predicts $\bar{y}$ for every input). $R_2 = 1$ means perfect prediction; all variance is explained by the model.

(c)

This is a supervised learning problem because the training data set has label for all the data points. We know how much the `fare_amount` is for each of the taxi rides in the training set.

(d)

This is a regression problem because want to learn a model that predicts continuous numerical values for the target.

Problem 2

$n = 1000$	Predicted Low-risk	Predicted High-risk
Actual Low-risk	TN = 850	FP = 50
Actual High-risk	FN = 20	TP = 80

(a)

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{80}{80 + 20} = 0.8 \tag{5}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{50}{50 + 850} = 0.0556 \tag{6}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{850}{850 + 50} = 0.9444 \tag{7}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} = \frac{20}{20 + 80} = 0.2 \tag{8}$$

(b)

In general, the cost of making a mistake depends on the type of the mistake. In the case of predicting risk level of borrowers, making a false positive mistake means that the company may refuse to give loan to a low-risk customer. Although this is undesirable, it does not incur in a significant cost to the company. On the other hand, making a false negative mistake can lead the company to giving loan to a high-risk customer, who is likely to default on their loan. So making a false negative mistake may result in a much higher cost than making a false positive one.

(c)

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} = \frac{850 + 80}{1000} = 0.93 \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{80}{80 + 50} = 0.615 \tag{10}$$

$$\text{Recall} = \text{TPR} = 0.8 \tag{11}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.615 \times 0.8}{0.615 + 0.8} = 0.695 \tag{12}$$

Problem 3

The loss function of L_2-regularized linear regression method is given by

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - \mathbf{w}^T x^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2 \\
&= \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2,
\end{aligned} \tag{13}$$

where w_0 is the intercept and $x_0^{(i)} = 1$.

For $j = 0$, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_0} &= \frac{1}{2}\sum_{i=1}^{N} 2\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)\frac{\partial}{\partial w_0}\left[-\sum_{j=0}^{D} w_j x_j^{(i)}\right] + 0 \\
&= \sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)(-x_0^{(i)}) \\
&= -\sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)
\end{aligned} \tag{14}$$

For $j = 1, 2, \ldots, D$, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_j} &= \frac{1}{2}\sum_{i=1}^{N} 2\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)\frac{\partial}{\partial w_j}\left[-\sum_{j=0}^{D} w_j x_j^{(i)}\right] + \frac{\lambda}{2}\frac{\partial}{\partial w_j}\left[\sum_{j=1}^{D} w_j^2\right] \\
&= \sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right)(-x_j^{(i)}) + \lambda w_j \\
&= -\sum_{i=1}^{N}\left(y^{(i)} - \sum_{j=0}^{D} w_j x_j^{(i)}\right) x_j^{(i)} + \lambda w_j
\end{aligned} \tag{15}$$

Problem 4

The loss function of L_2-regularized logistic regression method is given by

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln\left(\sigma(\mathbf{w}^T x^{(i)})\right) + (1-y^{(i)})\ln\left(1-\sigma(\mathbf{w}^T x^{(i)})\right)\right] + \frac{\lambda}{2N}\sum_{j=1}^{N} w_j^2 \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}\ln\left(\sigma(z^{(i)})\right) + (1-y^{(i)})\ln\left(1-\sigma(z^{(i)})\right)\right] + \frac{\lambda}{2N}\sum_{j=1}^{N} w_j^2 \\
&\equiv \mathcal{L}_c(\mathbf{w}) + \mathcal{L}_r(\mathbf{w}), \qquad (16)
\end{aligned}$$

where

$$z^{(i)} \equiv \mathbf{w}^T x^{(i)}, \qquad (17)$$

and

$$\sigma(z) = \frac{1}{1+e^{-z}}. \qquad (18)$$

For $j = 0$, we have

$$\frac{\partial \mathcal{L}_r(\mathbf{w})}{\partial w_0} = 0. \qquad (19)$$

For $j = 1, 2, \ldots, D$

$$\frac{\partial \mathcal{L}_r(\mathbf{w})}{\partial w_j} = \frac{\lambda}{N} w_j. \qquad (20)$$

For $j = 0, 1, 2 \ldots, D$

$$\begin{aligned}
\frac{\partial \mathcal{L}_c(\mathbf{w})}{\partial w_j} &= -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y^{(i)}}{\sigma(z^{(i)})}\frac{\partial \sigma(z^{(i)})}{\partial w_j} - \frac{1-y^{(i)}}{1-\sigma(z^{(i)})}\frac{\partial \sigma(z^{(i)})}{\partial w_j}\right] \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y^{(i)}}{\sigma(z^{(i)})} - \frac{1-y^{(i)}}{1-\sigma(z^{(i)})}\right]\frac{\partial \sigma(z^{(i)})}{\partial w_j}. \qquad (21)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \sigma(z^{(i)})}{\partial w_j} &= \frac{d\sigma(z^{(i)})}{dz^{(i)}}\frac{\partial z^{(i)}}{\partial w_j} \\
&= \frac{e^{-z^{(i)}}}{(1+e^{-z^{(i)}})^2} x_j^{(i)} \\
&= \left(\frac{1}{1+e^{-z^{(i)}}}\right)\left(\frac{1+e^{-z^{(i)}}-1}{1+e^{-z^{(i)}}}\right) x_j^{(i)} \\
&= \left(\frac{1}{1+e^{-z^{(i)}}}\right)\left(1-\frac{1}{1+e^{-z^{(i)}}}\right) x_j^{(i)} \\
&= \sigma(z^{(i)})(1-\sigma(z^{(i)}))x_j^{(i)} \qquad (22)
\end{aligned}$$

Substituting Eq. (22) into Eq. (21), we get

$$\begin{aligned}
\frac{\partial \mathcal{L}_c(\mathbf{w})}{\partial w_j} &= -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y^{(i)}}{\sigma(z^{(i)})} - \frac{1-y^{(i)}}{1-\sigma(z^{(i)})}\right]\sigma(z^{(i)})(1-\sigma(z^{(i)}))x_j^{(i)} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}(1-\sigma(z^{(i)})) - (1-y^{(i)})\sigma(z^{(i)})\right]x_j^{(i)} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} - \sigma(z^{(i)})\right]x_j^{(i)} \qquad (23)
\end{aligned}$$

So for $j = 0$, we have

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_0} = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} - \sigma(z^{(i)})\right]. \qquad (24)$$

For $j = 1, 2, \ldots, D$, we have

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)} - \sigma(z^{(i)})\right]x_j^{(i)} + \frac{\lambda}{N}w_j. \qquad (25)$$