National University of Singapore
School of Computing
CS3244: Machine Learning
Solution to Tutorial 3

Linear Classifiers and Logistic Regression

Colab Notebook Solutions: Linear Classifiers and Logistic Regression

1. **Logistic Regression**.

   Which of the following evaluation metrics cannot be applied in case of logistic regression output to compare with target? Explain your answer.

   (a) Accuracy
   (b) AUC-ROC
   (c) Log loss
   (d) Mean-Squared-Error

   Mean-Squared Error.

   - *Accuracy is a useful metric (Everyone knows this.)*
   - *The AUC-ROC curve is a metric tells how well the classifier is able to separate positive classes from negative classes.*
   - *We have already seen log loss in the class (We will see this in Question 3 of the tutorial). We want the loss to be as minimum as possible. It is a good measure of how good your learning algorithm is doing.*
   - *The mean squared error is used in regression problems. In case of logistic regression the true labels are discrete and it does not give us a good measure of success for learning.*

2. **Linear Regression Model Fitting**.

   You are given several data points as followed. Find a linear regression model that fits the data points best in terms of goodness-of-fit.

x_1	x_2	x_3	y
6	4	11	20
8	5	15	30
12	9	25	50
2	1	3	7

   We can just make use of the Normal Equation to solve for the weights θ.Extra thinking: Normal Equation needs the calculation of $(X^T X)^{-1}$. But sometimes this matrix is not invertible, when will that happen, and what should we do at that situation?

$$X = \begin{bmatrix} 1 & 6 & 4 & 11 \\ 1 & 8 & 5 & 15 \\ 1 & 12 & 9 & 25 \\ 1 & 2 & 1 & 3 \end{bmatrix}, Y = \begin{bmatrix} 20 \\ 30 \\ 50 \\ 7 \end{bmatrix}$$

$$\begin{aligned}\theta &= (X^T X)^{-1} X^T Y \\ &= \begin{bmatrix} 4 & -5.5 & -7 & 7 \end{bmatrix}^T, \\ \hat{y} &= 4 - 5.5x_1 - 7x_2 + 7x_3\end{aligned}$$

3. **Gradient of the Logistic Regression Cost Function**

a) Derivative of the sigmoid function. *The sigmoid function is given below. Express the derivative of the sigmoid function with respect to z in terms of the sigmoid function*

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

$$\begin{aligned}\frac{\partial \sigma(z)}{\partial z} &= -\frac{1}{(1+e^{-z})^2}.\frac{\partial(1+e^{-z})}{\partial z} \\ &= \frac{e^{-z}}{(1+e^{-z})^2} \\ &= \frac{1}{(1+e^{-z})}.\frac{e^{-z}}{(1+e^{-z})} \\ &= \frac{1}{(1+e^{-z})}.\frac{1+e^{-z}-1}{(1+e^{-z})} \\ &= \sigma(z)\{\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\} \\ &= \sigma(z)(1-\sigma(z))\end{aligned}$$

b) Derivative of the sigmoid function continued *- What is the derivative of the function $log(1-\sigma(z))$ with respect to z.*

$$\begin{aligned}h(x) &= log(1-\sigma(z)) \\ &= -\frac{1}{1-\sigma(z)}\frac{\partial \sigma(z)}{\partial z} \\ &= -\frac{\sigma(z)(1-\sigma(z))}{(1-\sigma(z))} \\ &= -\sigma(z)\end{aligned}$$

c) Derivative of the cost function *Given data points $\{(x_1, y_1), (x_2, y_2)....(x_m, y_m)\}$, the cost function for logistic regression is given by*

$$L(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y_i \ log(h_\theta(x_i)) + (1-y_i) \ log(1-h_\theta(x_i))$$

where $h_\theta(x) = \sigma(\theta^T x)$. Derive the gradient of the cost function with respect to the parameters θ.

$$
\begin{aligned}
p(x_i) &= y_i \;\; log(\sigma(\theta^T x_i)) \\
\Rightarrow \frac{\partial(p(x_i))}{\partial\theta} &= y_i \cdot \frac{1}{\sigma(\theta^T \cdot x_i)} \cdot \frac{\partial\sigma(\theta^T \cdot x_i)}{\partial\theta} \\
&= y_i \cdot \frac{1}{\sigma(\theta^T \cdot x_i)} \cdot \sigma(\theta^T \cdot x_i)(1 - \sigma(\theta^T \cdot x_i)) \cdot \frac{\partial(\theta^T \cdot x_i)}{\partial\theta} \\
&= y_i \cdot \frac{1}{\sigma(\theta^T \cdot x_i)} \cdot \sigma(\theta^T \cdot x_i)(1 - \sigma(\theta^T \cdot x_i)) \cdot x_i \\
&= x_i \cdot y_i(1 - \sigma(\theta^T \cdot x_i))
\end{aligned}
$$

$$
\begin{aligned}
q(x_i) &= (1 - y_i) log(1 - h_\theta(x_i)) \\
\frac{\partial q(x_i)}{\partial\theta} &= (1 - y_i)\frac{\partial log(1 - \sigma(\theta^T x_i))}{\partial\theta} \\
&= (1 - y_i)\sigma(\theta^T x_i)(-x_i)
\end{aligned}
$$

Add the above two formulae -

$$
\frac{\partial L}{\partial(\theta)} = -\frac{1}{m}\sum_{i=1}^{m} x_i(y_i - \sigma(\theta^T . x_i)) \tag{1}
$$

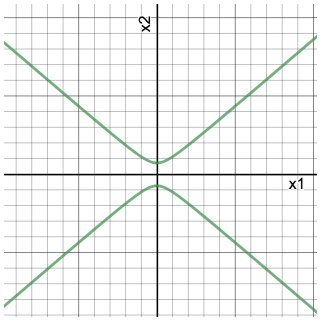
4. **Nonlinear Transformations.**

Consider the feature transform $\Phi(1, x_1, x_2) = (1, x_1^2, x_2^2)$. Draw and show the boundary (not strictly) in $\mathcal{X}$ that a hyperplane $\hat{\theta}$ in $\mathcal{Z}$ correspond to under the following cases:

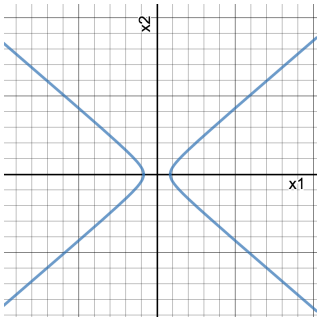
check out the webcast for further details

(a) $\hat{\theta}_1 > 0, \hat{\theta}_2 < 0$.

When $\hat{\theta}_0 > 0$, the plot shows a "North-South opening hyperbola"; when $\hat{\theta}_0 < 0$, it's a "East-West opening hyperbola" as shown in Figure 1.



(a) Borders when $\hat{\theta}_0 > 0$



(b) Borders when $\hat{\theta}_0 < 0$

Figure 1: Borders for the two cases of $\hat{\theta}_0$

(b) $\hat{\theta}_1 > 0, \hat{\theta}_2 = 0$.
When $\hat{\theta}_0 > 0$*, there's no border in* $\mathcal{X}$ *space, all data points* $\mathbf{x}$ *are classified into* $y = +1$ *category. When* $\hat{\theta}_0 < 0$*, the border is a single point at* x_1 *axis as shown in Figure 2.*

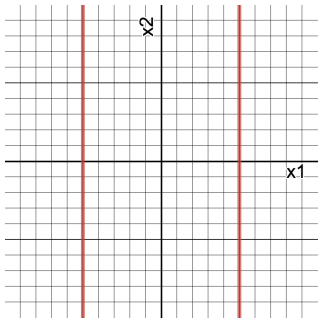


Figure 2: Borders when $\hat{\theta}_0 < 0$

(c) $\hat{\theta}_1 > 0, \hat{\theta}_2 > 0, \hat{\theta}_0 < 0$.
It's either a ellipse or a circle, as shown in Figure 3

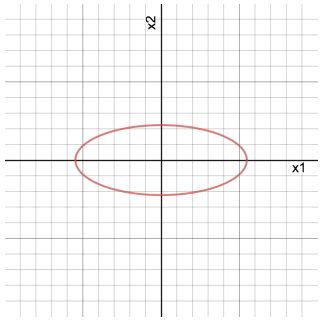


Figure 3: Borders when $\hat{\theta}_0 < 0$

(d) $\hat{\theta}_1 > 0, \hat{\theta}_2 > 0, \hat{\theta}_0 > 0$.
There's no border in $\mathcal{X}$ *space. All data points* $\mathbf{x}$ *are classified into* $y = +1$ *category as well.*

5. **[**]The Hat Matrix.**

The hat matrix is an integral part of understanding linear regression. We see the hat matrix makes an appearance in the analytical, closed form solution of linear regression, as it contains the pseudo inverse $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$*. Let's look at its properties.*

Consider the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$*, where* $\mathbf{X}$ *is an* m *by* $n+1$ *matrix, and* $\mathbf{X}^\top\mathbf{X}$ *is invertible.*

(a) Show that $\mathbf{H}$ *is symmetric.*

Show $\mathbf{H}^\top = \mathbf{H}$.

$$\begin{aligned}
\mathbf{H}^\top &= (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top)^\top && \textit{Expand} && (2)\\
&= (\mathbf{X}^\top)^\top(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}})^\top && \textit{Use } (\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top && (3)\\
&= \mathbf{X}((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}})^\top\mathbf{X}^\top && \textit{Use } (\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top && (4)\\
&= \mathbf{X}((\mathbf{X}^\top\mathbf{X})^\top)^{-\mathbf{1}})\mathbf{X}^\top && \textit{Use } (\mathbf{A}^\top)^{-\mathbf{1}} = (\mathbf{A}^{-\mathbf{1}})^\top && (5)\\
&= \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top. && && (6)\\
&= \mathbf{H}. && \square && (7)
\end{aligned}$$

(b) Show that $\mathbf{H}^k = \mathbf{H}$*, for any positive integer* k.

Base Case ($k = 1$)*:* $\mathbf{LHS} = \mathbf{H^1} = \mathbf{H} = \mathbf{RHS}$. *Hence, base case is true.*
Inductive Step: Assume $H^k = H$ *for any* $k \geq 1$*, show for* $k+1$. *Show* $\mathbf{H^{k+1}} = \mathbf{H}$.

$$\begin{aligned}
\mathbf{H^{k+1}} &= (\mathbf{H^k})(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) && \textit{Expand} && (8)\\
\mathbf{H^{k+1}} &= (\mathbf{H})(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) && \textit{Premise} && (9)\\
&= (\mathbf{X}\underbrace{(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}}_{A}\underbrace{(\mathbf{X}^\top\mathbf{X})}_{B}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) && \textit{Associate} && (10)\\
&= (\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top) && \textit{Cancel } A \textit{ \& } B \textit{ to } \mathbf{I}\textit{, drop } \mathbf{I} && (11)\\
&= \mathbf{H}. && \square && (12)
\end{aligned}$$

This property $\mathbf{X} = \mathbf{X}^k$ *is called* **idempotency**. *In fact, orthogonal projection matrices (which* $\mathbf{H}$ *is a member of), exactly have both the symmetric and idempotent properties (i.e., "if and only if")*

(c) If $\mathbf{I}$ *is the identity matrix of size* m*, show that* $(\mathbf{I}-\mathbf{H})^{\mathrm{K}} = \mathbf{I}-\mathbf{H}$ *for any positive integer* K.

We can prove this similar to Part (b), using an inductive proof. We just detail the core part here.

$$\begin{aligned}
(\mathbf{I}-\mathbf{H})^{\mathbf{2}} &= (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) && \textit{Expand} && (1)\\
&= \mathbf{I^2} - \mathbf{2IH} + \mathbf{H^2} && \textit{Multiply out} && (2)\\
&= \mathbf{I} - \mathbf{2IH} + \mathbf{H} && \mathbf{I} \textit{ and } \mathbf{H}'\textit{s idempotency; Part } (b) && (3)\\
&= \mathbf{I} - \mathbf{H}. && \textit{Drop } \mathbf{I}, \textit{associate } \square && (4)
\end{aligned}$$

You might guess that $(\mathbf{I}-\mathbf{H})$ *also is an orthogonal projection matrix, since it is also symmetric and idempotent. And you'd be right.*

(d) Show that $trace(\mathbf{H}) = n+1$*, where the trace (tr) is the sum of diagonal elements. [Hint:* $trace(\mathbf{AB}) = trace(\mathbf{BA})$*.]*

$$
\begin{aligned}
trace(\mathbf{H}) &= trace(\underbrace{\mathbf{X}}_{\mathbf{A}}\underbrace{(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top}_{\mathbf{B}}) && \textit{Expand} && (1)\\
&= trace((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}) && \textit{Apply hint} && (2)\\
&= trace(\mathbf{I}_{(n+1)}). && \underbrace{\mathbf{X}^{-\mathbf{1}}}_{(n+1)\times m}\ \underbrace{\mathbf{X}}_{m\times(n+1)} = \mathbf{I} && (3)
\end{aligned}
$$

For a symmetric and idempotent matrix A*,* $rank(A) = trace(A)$*. We'll use these properties in the next question.*