CS 6140: Machine Learning — Fall 2021— Paul Hand

Midterm 1 Study Guide and Practice Problems
Due: Never.

Names: Sample Solutions

This document contains practice problems for Midterm 1. The midterm will only have 5 problems. The midterm will cover material up through and including the bias-variance tradeoff, but not including ridge regression. Skills that may be helpful for successful performance on the midterm include:

1. Setting up and solving a linear regression problem with features that are nonlinear functions of the model's input.
2. Writing down the optimization problem for least squares linear regression using matrix-vector notation
3. Familiarity with matrix multiplication, in particular when multiplying by diagonal matrices
4. Evaluating the true positive rate, false positive rate, precision, and recall of a predictor for binary classification
5. Setting up a logistic regression problem and writing down the appropriate function that is being minimized
6. Computing the mean, expected value, and variance of uniform random variables
7. Explaining causes and remedies for overfitting and underfitting of ML models

Question 1.

Consider the following training data.

x_1	x_2	y
0	0	0
0	1	1.5
1	0	2
1	1	2.5

Suppose the data comes from a model $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$+noise for unknown constants $\theta_0, \theta_1, \theta_2$. Use least squares linear regression to find an estimate of $\theta_0, \theta_1, \theta_2$.

Response:

Let $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 0 \\ 1.5 \\ 2 \\ 2.5 \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$

We solve the least squares problem

$$\min_{\theta} \| y - X\theta \|^2$$

The solution is given by

$$\theta = (X^t X)^{-1} X^t y$$

Solving $X^t X = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}$, $(X^t X)^{-1} = \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$

So $\theta = \begin{pmatrix} 0.5 \\ 2 \\ 0.5 \end{pmatrix} \Rightarrow \boxed{\begin{aligned} \theta_0 &= 0.5 \\ \theta_1 &= 2 \\ \theta_2 &= 0.5 \end{aligned}}$

Question 2.

Consider the following training data:

x	y
1	3
2	1
3	0.5

Suppose the data comes from a model $y = cx^{\beta}$ + noise, for unknown constants c and β. Use least squares linear regression to find an estimate of c and β.

Response:

Write $\log y = \underbrace{\log c}_{\theta_1} + \underbrace{\beta}_{\theta_2} \log x$

Data becomes:

$\log x$	$\log y$
0	$\log 3$
$\log 2$	0
$\log 3$	$\log 0.5$

Let $\underset{\sim}{X} = \begin{pmatrix} 1 & 0 \\ 1 & \log 2 \\ 1 & \log 3 \end{pmatrix}$, $y = \begin{pmatrix} \log 3 \\ 0 \\ \log 0.5 \end{pmatrix}$, $\theta = \begin{pmatrix} \log c \\ \beta \end{pmatrix}$

Solve $\min_{\theta} \| y - \underset{\sim}{X}\theta \|^2$

Solution given by $\theta = (X^t X)^{-1} X^t y$

Using numpy, we compute $\theta = \begin{pmatrix} 0.899 \\ -0.5 \end{pmatrix}$

$\Rightarrow$ $c = e^{0.899}$
$\beta = -0.5$

$\Rightarrow$ $\boxed{\begin{array}{l} c = 2.45 \\ \beta = -0.5 \end{array}}$

Question 3.

(a) Let $\theta^* \in \mathbb{R}^d$, and let $f(\theta) = \frac{1}{2}\|\theta - \theta^*\|^2$. Show that the Hessian of f is the identity matrix.

Response:

(b) Let $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$. For $\theta \in \mathbb{R}^d$, let $g(\theta) = \frac{1}{2}\|X\theta - y\|^2$. Show that the Hessian of g is $X^t X$.

Response:

a) $$(H)_{jk} = \frac{\partial^2}{\partial\theta_j \partial\theta_k} f(\theta)$$

Write $$f(\theta) = \frac{1}{2}\sum_{i=1}^{d}(\theta_i - \theta_i^*)^2$$

$$\frac{\partial}{\partial\theta_k} f(\theta) = (\theta_k - \theta_k^*)$$

$$\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_k} f(\theta) = \frac{\partial}{\partial\theta_j}(\theta_k - \theta_k^*) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

So $$H_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}. \Rightarrow \boxed{H = I_d}$$

b) $g(\theta) = \frac{1}{2} \|X\theta - y\|^2$

By class,

$$\nabla g(\theta) = X^t(X\theta - y)$$
$$= X^t X\theta - X^t y$$

Let $M = X^t X \in \mathbb{R}^{d \times d}$. Let m_k be k^{th} row of M

So $$\frac{\partial g}{\partial \theta_k} = (M\theta - X^t y)_k$$
$$= m_k^t \theta - (X^t y)_k$$

So $$\frac{\partial}{\partial \theta_j} \frac{\partial g}{\partial \theta_k} = m_{kj}$$

j^{th} entry of m_k.

Thus $$\boxed{H = M = X^t X.}$$

Question 4.

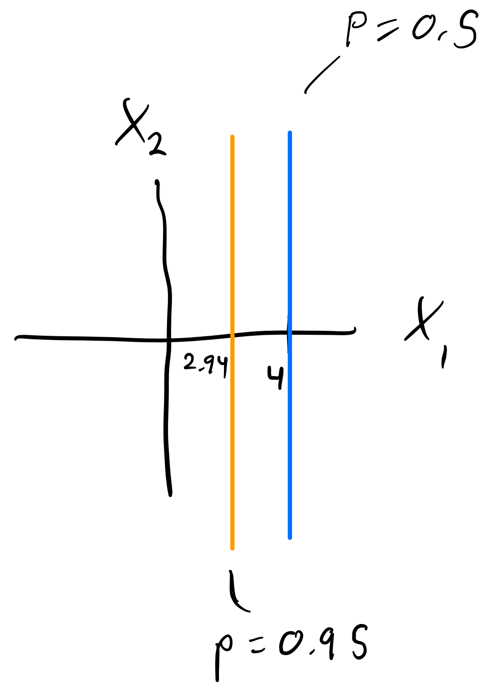
Consider a binary classification problem whose features are in $\mathbb{R}^2$. Suppose the predictor learned by logistic regression is $\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$, where $\theta_0 = 4, \theta_1 = -1, \theta_2 = 0$. Find and plot curve along which P(class 1) = 1/2 and the curve along which P(class 1) = 0.95.

Response:

$$\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = 1/2$$

$$\Rightarrow \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

$$\Rightarrow \quad 4 - x_1 = 0$$



$$\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = 0.95$$

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

So $\sigma(z) = 0.95 \Rightarrow \dfrac{1}{1+e^{-z}} = 0.95$

$$\Rightarrow e^{-z} = -1 + \frac{1}{0.95} = 0.0526$$

$$\Rightarrow z = 2.94$$

Question 5.

Consider a 3-class classification problem. You have trained a predictor whose input is $x \in \mathbb{R}^2$ and whose output is softmax$(x_1 + x_2 - 1, 2x_1 + 3, x_2)$. Find and sketch the three regions in $\mathbb{R}^2$ that gets classified as class 1, 2, and 3.

Response:

The predicted class corresponds to the largest component of softmax, which is the same as the largest input to softmax.

$$Z_1 = X_1 + X_2 - 1$$

$$Z_2 = 2X_1 + 3$$

$$Z_3 = X_2$$

a) where is classified as class 1?

$$X_1 + X_2 - 1 > 2X_1 + 3 \quad \& \quad X_1 + X_2 - 1 > X_2$$

$$\Rightarrow \quad -X_1 + X_2 > 4 \quad \& \quad X_1 > 1$$

b) Where is classified as class 2 :

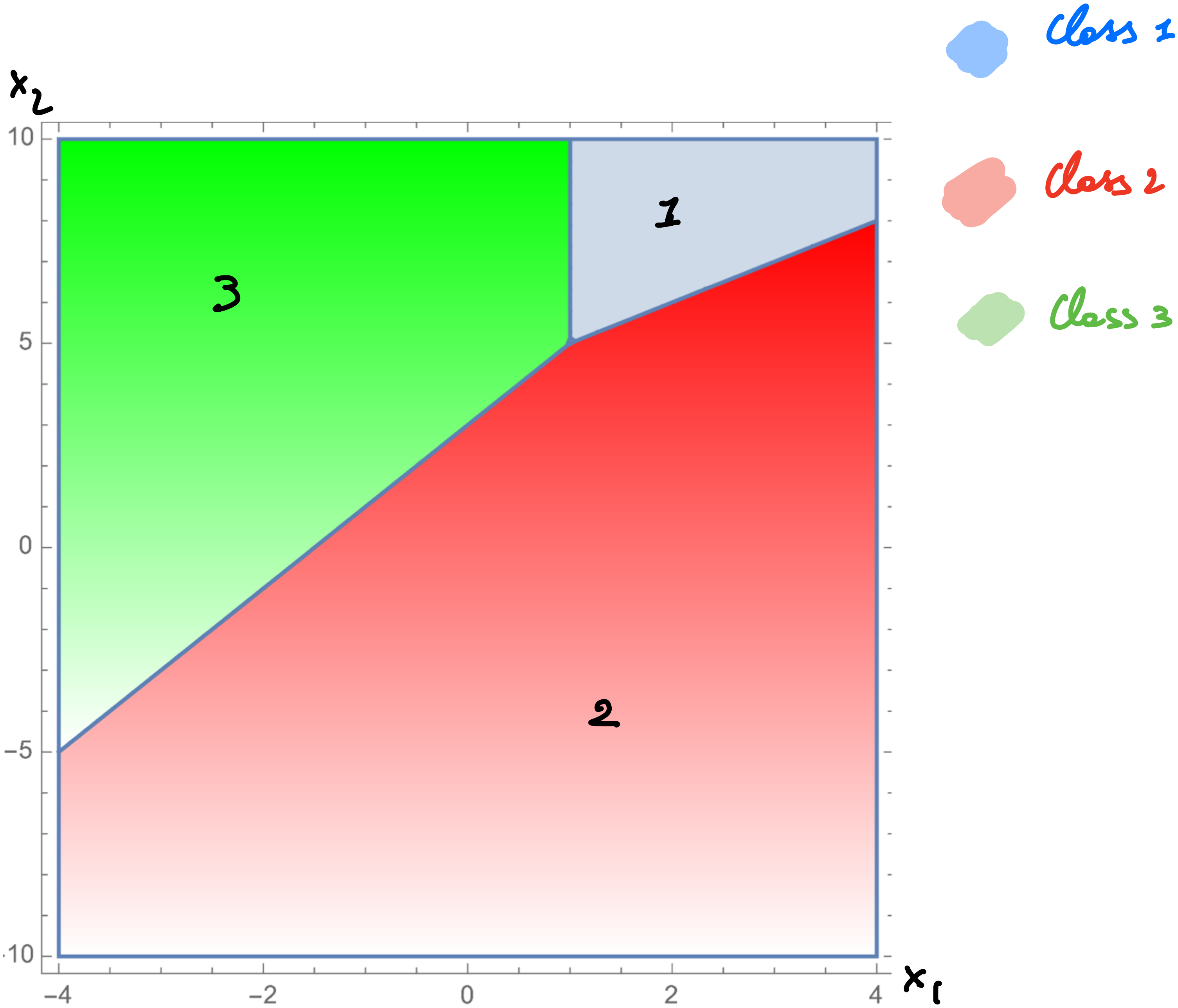
$2X_1 + 3 > X_1 + X_2 - 1$ & $2X_1 + 3 > X_2$

$-X_1 + X_2 < 4$ & $2X_1 - X_2 > -3$

c) Where is classified as class 3 :

$X_2 > X_1 + X_2 - 1$ & $X_2 > 2X_1 + 3$

$X_1 < 1$ & $2X_1 - X_2 < -3$



x_2
10
5
0
−5
-10
−4
−2
0
2
4
x_1
1
2
3
Class 1
Class 2
Class 3

Question 6.

Suppose $x \sim$Uniform$([-1,1])$ and $y = x + \varepsilon$, where $\varepsilon \sim$Uniform$([-\gamma,\gamma])$ for some $\gamma > 0$. Consider a predictor given by $f_\theta(x) = \theta_1 + \theta_2 x$, where $\theta \in \mathbb{R}^2$. Evaluate the risk of f_θ with respect to the square loss. Your answer should be a deterministic expression only depending on θ_1, θ_2, and γ.

Response:

$$R(\theta) = \mathbb{E}_{x,\varepsilon}\left[\left|f_\theta(x) - y\right|^2\right] = \mathbb{E}_{x,\varepsilon}\left[\left(\theta_1 + \theta_2 x - x - \varepsilon\right)^2\right]$$

$$= \mathbb{E}_{x\varepsilon}\left[\left(\theta_1 + (\theta_2 - 1)x - \varepsilon\right)^2\right]$$

Note: $\mathbb{E}[x^2]$

$= \int_{-1}^{1} x^2 \frac{1}{2} dx = \frac{1}{6} x^3 \Big|_{-1}^{1}$

$= 1/3$

$$= \mathbb{E}_{x\varepsilon}[\theta_1^2] + \mathbb{E}_{x\varepsilon}\left[(\theta_2-1)^2 x^2\right] + \mathbb{E}[\varepsilon^2]$$

$$-2\,\cancel{\mathbb{E}_{x\varepsilon}[\theta_1 \varepsilon]}^{\,0} - 2\,\cancel{\mathbb{E}_{x\varepsilon}[\theta_1(\theta_2-1)x]}^{\,0} - 2\,\cancel{\mathbb{E}_{x\varepsilon}[(\theta_2-1)x\varepsilon]}^{\,0}$$

Since θ_1, θ_2 are deterministic $\Rightarrow \mathbb{E}[\theta_1 \varepsilon] = \theta_1 \mathbb{E}[\varepsilon]$

$\mathbb{E}[x] = \mathbb{E}[\varepsilon] = 0$ and x and ε indep. $\Rightarrow \mathbb{E}[x\varepsilon] = \mathbb{E}[x]\,\mathbb{E}[\varepsilon]$

$$R(\theta) = \theta_1^2 + \frac{1}{3}(\theta_2 - 1)^2 + \frac{1}{3}\gamma^2$$

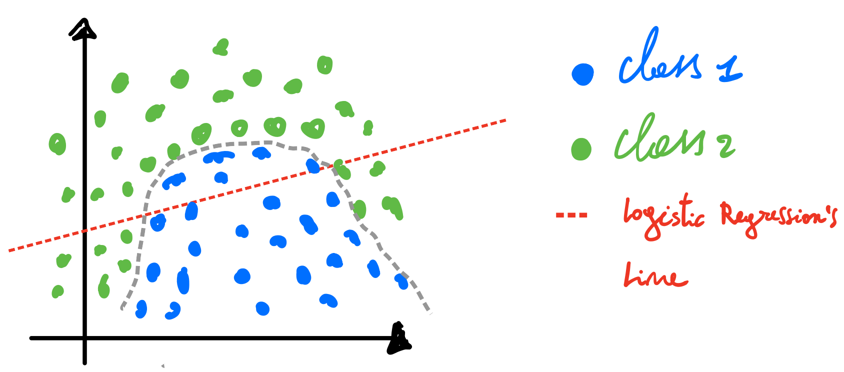
Question 7.

You are training a logistic regression model and you notice that it does not perform well on test data.

- Could the poor performance be due to underfitting? Explain.
- Could the poor performance be due to overfitting? Explain.

Underfitting Yes, logist regression separates the 2 classes using only a line. this might be too simple to explain the variations in the data

Consider for example the case of 2 classes separable by a curved line.



Overfitting

Yes, if there are too many features, the data could appear to be linearly separable as a mathematical artifact. This could result in overfitting of training data.