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

Linear Regression

<u>Data</u> $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ $x \in \mathbb{R}^d$

<u>Model</u> $y = f_\theta(x) + \text{noise}$

means find θ s.t.

$$y^{(i)} \approx f_\theta(x^{(i)})$$

Linear Regression

$$f_\theta(x) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_d$$

Least Squares

$$x^{(i)} \in \mathbb{R}^{d}, \qquad \theta \in \mathbb{R}^{d+1}, \qquad y^{(i)} \in \mathbb{R}$$

$$X = \begin{bmatrix} 1, & — & x^{(1)T} & — \\ \vdots & & \vdots & \\ 1, & — & x^{(n)T} & — \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$$

if bias is present

$$y = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^{n}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \vdots \\ \theta_d \end{bmatrix} \in \mathbb{R}^{d+1}$$

Then

$$\min_{\theta} \frac{1}{2} \| y - X\theta \|_2^2$$

$$= \min_{\theta} \frac{1}{2} \sum_{i=1}^{n} \left[y^{(i)} - \left(\theta_0 + \theta_1 x_1^{(i)} + \dots + \theta_d x_d^{(i)} \right) \right]^2$$

<u>Solution</u>

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

(if X has rank d)

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 $\theta = \begin{pmatrix} 0.5 \\ 2 \\ 0.5 \end{pmatrix} \Rightarrow \boxed{\begin{array}{l} \theta_0 = 0.5 \\ \theta_1 = 2 \\ \theta_2 = 0.5 \end{array}}$

Remark

Sometimes a nonlinear model model can be turned to linear model via some nonlinear transformation of the data

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

$f : \mathbb{R}^d \to \mathbb{R}$

gradient

$\nabla f(\theta) \in \mathbb{R}^{d \times 1}$

$$[\nabla f(\theta)]_k = \frac{\partial}{\partial \theta_k} f(\theta) \qquad \begin{bmatrix} \partial_{\theta_1} f \\ \vdots \\ \partial_d f \end{bmatrix}$$

or

$$f(\theta + h) = f(\theta) + V^T h + o(\|h\|)$$

then $\nabla f(\theta) = V$

Hessian

$$\nabla^2 f(\theta) \text{ or } H(\theta) \in \mathbb{R}^{d \times d}$$

$$[H(\theta)]_{jk} = \frac{\partial}{\partial \theta_j} \frac{\partial}{\partial \theta_k} f(\theta) = \frac{\partial}{\partial \theta_k} \frac{\partial}{\partial \theta_j} f(\theta) = [H(\theta)]_{kj}$$

or

$$J \nabla f(\theta) = \begin{bmatrix} - \nabla_\theta (\partial_{\theta_1} f)^T - \\ \vdots \\ - \nabla_\theta (\partial_{\theta_d} f)^T - \end{bmatrix}$$

(Derivative / Jacobian of the gradient)

or

$$f(\theta + h) = f(\theta) + \nabla f(\theta)^T h + \frac{1}{2} h^T H(\theta) h + O(\|h\|^2)$$

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

Matrix Multiplication

- $A \in \mathbb{R}^{d_2 \times d_1}$, $B \in \mathbb{R}^{d_1 \times d_3}$

  $AB \in \mathbb{R}^{d_2 \times d_3}$

- If you find that for $A \in \mathbb{R}^{d_2 \times d_1}$

$$A = \underbrace{\text{something}}_{d_2 \times d_1} + \underbrace{\text{something else}}_{d_2 \times d_1}$$

- $A \in \mathbb{R}^{d \times d}$ (SQUARE!)

  and invertible $A^{-1}A = AA^{-1}$

- $A \in \mathbb{R}^{d \times d}$ diagonal matrix if

when $i \neq j$ $A_{ij} = 0$

Note you can still have $A_{ii} = 0$

- Note that if $x \in \mathbb{R}^d$ and A diagonal

$$[Ax]_i = \underline{A_{ii}}\, \underline{x_i} \qquad [Ax]_2 = A_{22}\, x_2$$

- $B \in \mathbb{R}^{d \times m}$ and $A \in \mathbb{R}^{d \times d}$ diagonal

if $B = \begin{bmatrix} - b_1^T - \\ \vdots \\ - b_d^T - \end{bmatrix}$ b_i i-th row

$$AB = \begin{bmatrix} - A_{11} b_1^T - \\ \vdots \\ - A_{dd} b_d^T - \end{bmatrix}$$

- $C \in \mathbb{R}^{m \times d} = \begin{bmatrix} c_1, \dots, c_d \end{bmatrix}$ $\quad c_i$ i-th column

$$CA = \begin{bmatrix} | & & | \\ A_{11} c_1 & , \dots , & A_{dd} c_D \\ | & & | \end{bmatrix}$$

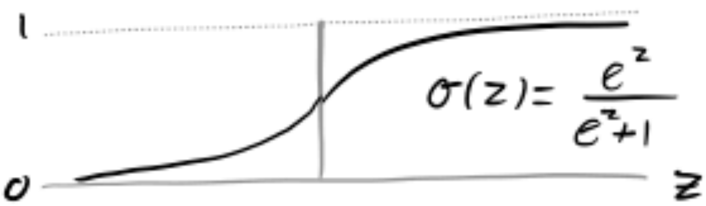
Binary Classification

Model $y = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = \hat{y}(x;\theta)$

where

"Sigmoid"

"logistic function"



Solve $\min\limits_{\theta} \sum_{i=1}^{n} L\left(y_i,\ \hat{y}(x^{(i)};\theta)\right)$ for $\hat{\theta}$

Predict:

For new sample x, predict

$$\begin{cases} \text{class } 1 & \text{if } \hat{y} \geq 1/2 \\ \text{class } 0 & \text{if } \hat{y} < 1/2 \end{cases}$$

Decision boundary $\hat{y} = \frac{1}{2}$

$\sigma(z) = \frac{1}{2}$ if $z = 0$

$\Rightarrow$ Decision boundary: $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

Question 4.

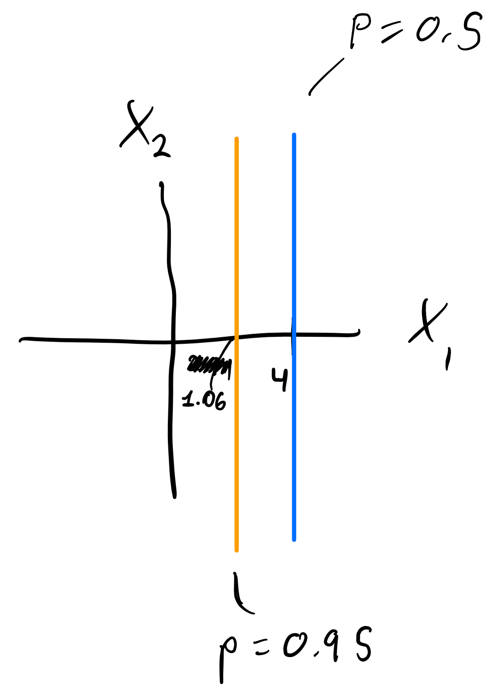
Consider a binary classification problem whose features are in $\mathbb{R}^2$. Suppose the predictor learned by logistic regression is $\sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$, where $\theta_0 = 4, \theta_1 = -1, \theta_2 = 0$. Find and plot curve along which P(class 1) = 1/2 and the curve along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

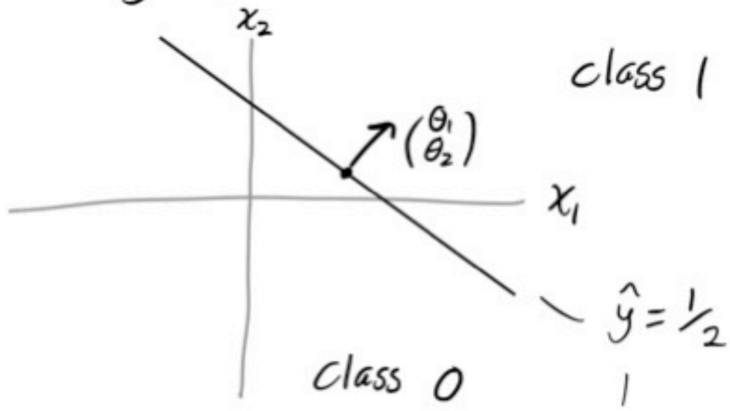
So $\quad \sigma(z) = 0.95 \Rightarrow \dfrac{1}{1+e^{-z}} = 0.95$

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

$$\Rightarrow \quad z = 2.94$$

$$\Rightarrow \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 2.94 \Rightarrow 4 - x_1 = 2.94 \Rightarrow x_1 = 1.06$$

Decision boundary is linear



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

Multiclass Classification

Consider a 3 class classification problem (w/ no bias term)

Training Data: $\{(x^{(i)}, y^{(i)})\}_{i=1\cdots n}$

$x^{(i)} \in \mathbb{R}^d$ for all i

$y^{(i)} \in \mathbb{R}^3$, $y^{(i)} = \begin{cases} \begin{pmatrix}1\\0\\0\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 1 \\ \begin{pmatrix}0\\1\\0\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 2 \\ \begin{pmatrix}0\\0\\1\end{pmatrix} & \text{if } y^{(i)} \text{ of class } 3 \end{cases}$

one-hot encoding

Model: $y = \text{softmax}(z_1, z_2, z_3)$

logits $\begin{cases} z_1 = \theta_1 \cdot x \\ z_2 = \theta_2 \cdot x \\ z_3 = \theta_3 \cdot x \end{cases}$ w/ $\theta_1 \in \mathbb{R}^d$, $\theta_2 \in \mathbb{R}^d$, $\theta_3 \in \mathbb{R}^d$

"logits are linear in parameters of model"

$$\text{softmax}(z_1, z_2, z_3) = \begin{pmatrix} \dfrac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}} \\ \dfrac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}} \\ \dfrac{e^{z_3}}{e^{z_1}+e^{z_2}+e^{z_3}} \end{pmatrix}$$

— adds up to 1

Train:

$$\min_{\theta} \sum_{i=1}^{n} L\left(y^{(i)}, \text{softmax}\left(\theta_1 \cdot x^{(i)}, \theta_2 \cdot x^{(i)}, \theta_3 \cdot x^{(i)}\right)\right)$$

log loss:

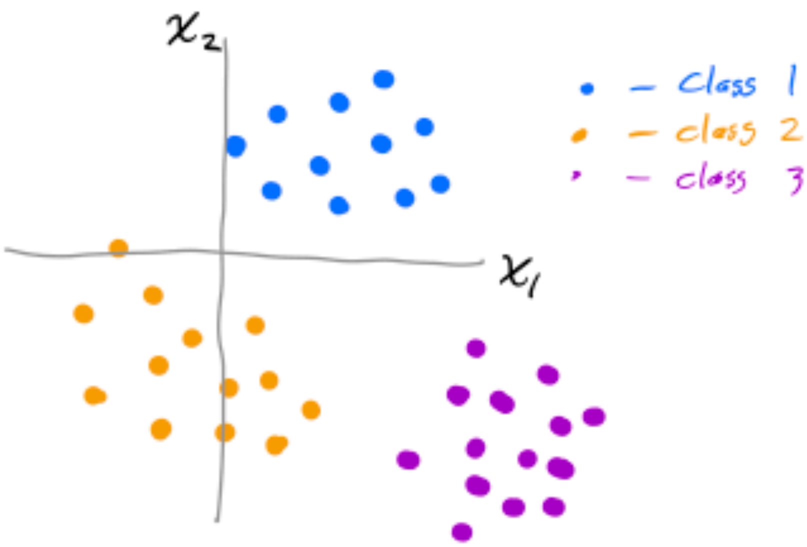
$$L(\underset{\mathbb{R}^3}{y}, \underset{\mathbb{R}^3}{\hat{y}}) = -\sum_{c=1}^{3} y_c \log \hat{y}_c$$

Prediction:

Given x, compute $\hat{y}(x; \hat{\theta}) \in \mathbb{R}^3$

Predict class $\hat{c}$ w/ $\hat{c} = \arg\max_{c} \hat{y}_c$

What will decision boundary look like for 3-class classification?



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 :

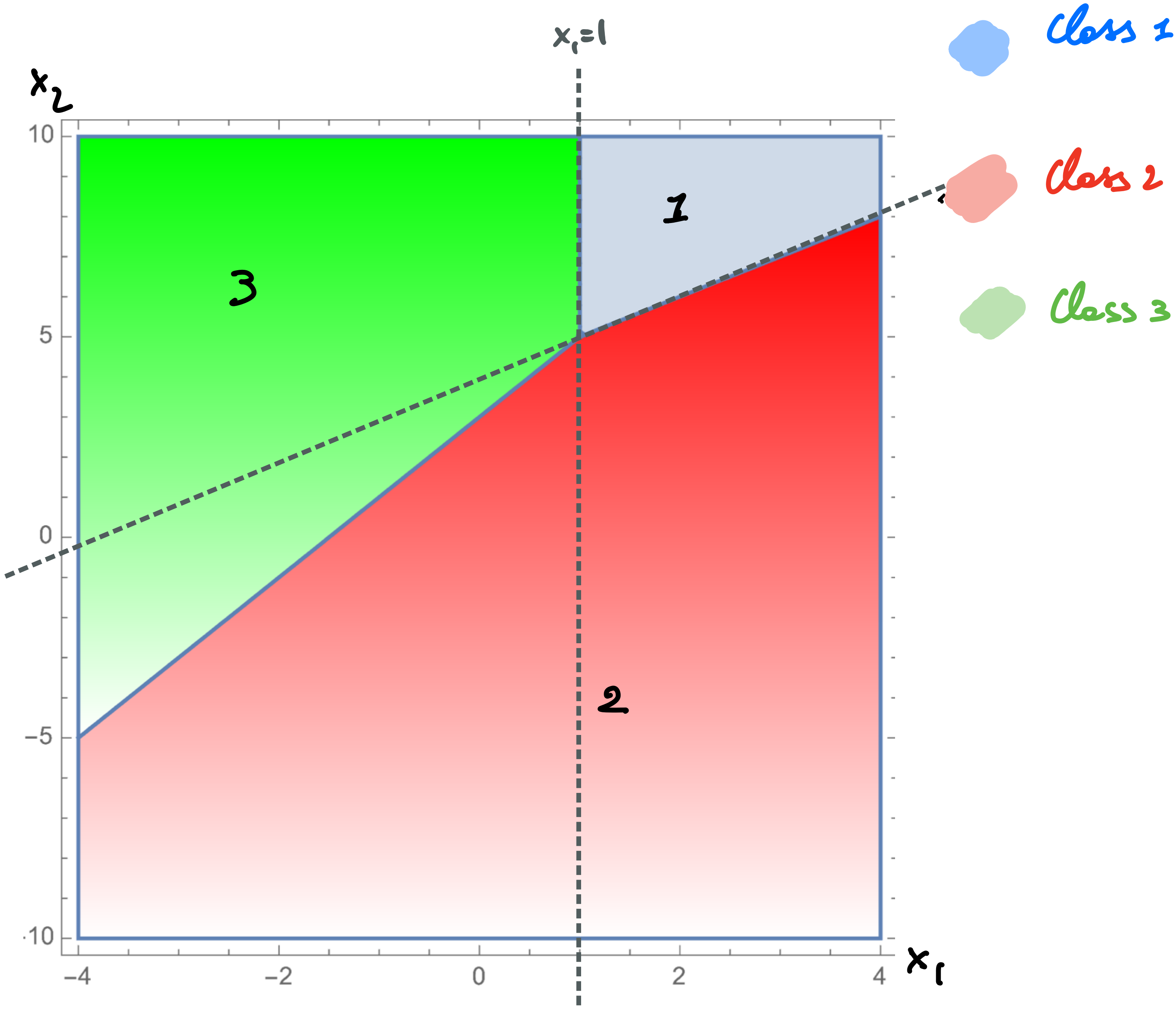
$$2X_1 + 3 > X_1 + X_2 - 1 \quad \& \quad 2X_1 + 3 > X_2$$

$$\underline{-X_1 + X_2 < 4} \quad \& \quad \underline{2X_1 - X_2 > -3}$$

c) Where is classified as class 3 :

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

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

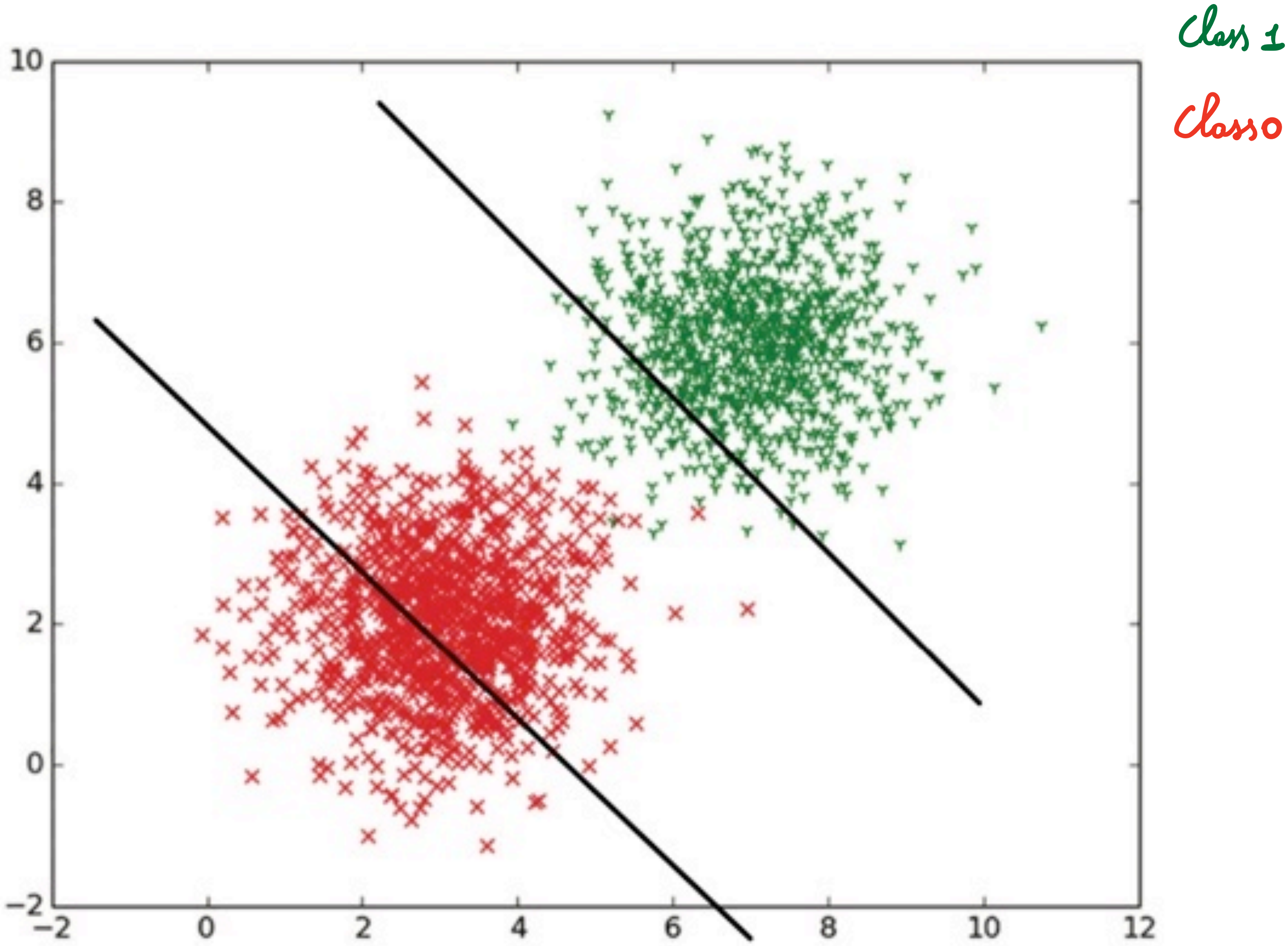


$x_1=1$
x_2
10
5
0
-5
-10
-4
-2
0
2
4
x_1
1
2
3
Class 1
Class 2
Class 3

Precision : $\frac{TP}{TP+FP}$ what fraction of predicted positives are real ?

Recall : $\frac{TP}{TP+FN}$ what fraction of positives are correctly predicted ?

Want high precision & high recall



RISK

- $f : x \mapsto f(x)$ predictor
- ℓ loss $\Rightarrow (\quad)^2$

Risk of f is

$$R(f) = \mathbb{E}_{(x,y)\sim D}\, \ell\big(y, f(x)\big)$$

If f depends on some θ then

$$R(\theta) = R(f_\theta) = \mathbb{E}_{(x,y)\sim D}\, \ell\big(y, f_\theta(x)\big)$$

Note θ here is just a parameter fixed (constant)

Note when we study Bias-Variance :

then $\hat{\theta}$ is estimated from data $\Rightarrow$ Random

so we also take expectation over

S = samples

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

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

$$= 1/3$$

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

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

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. $\Rightarrow \mathbb{E}[x\varepsilon] = \mathbb{E}[x]\,\mathbb{E}[\varepsilon]$

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

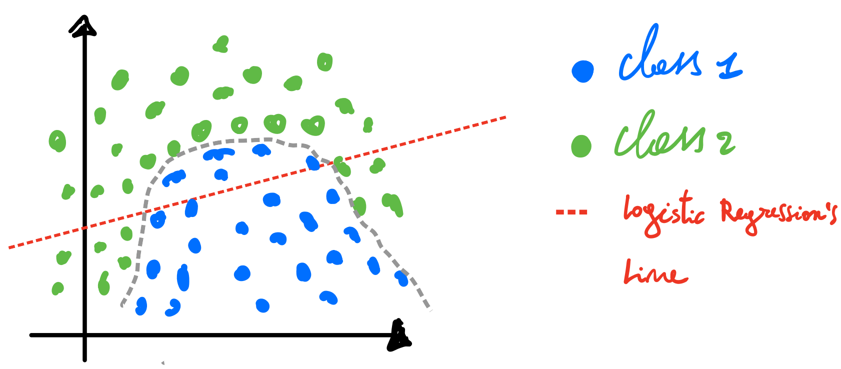
Question 7.

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.

<u>Underfitting</u> Yes, logist regression seperates 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 seperable 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.