01.112/50.007 Machine Learning

Homework 2

Support Vector Machines & Logistic Regression

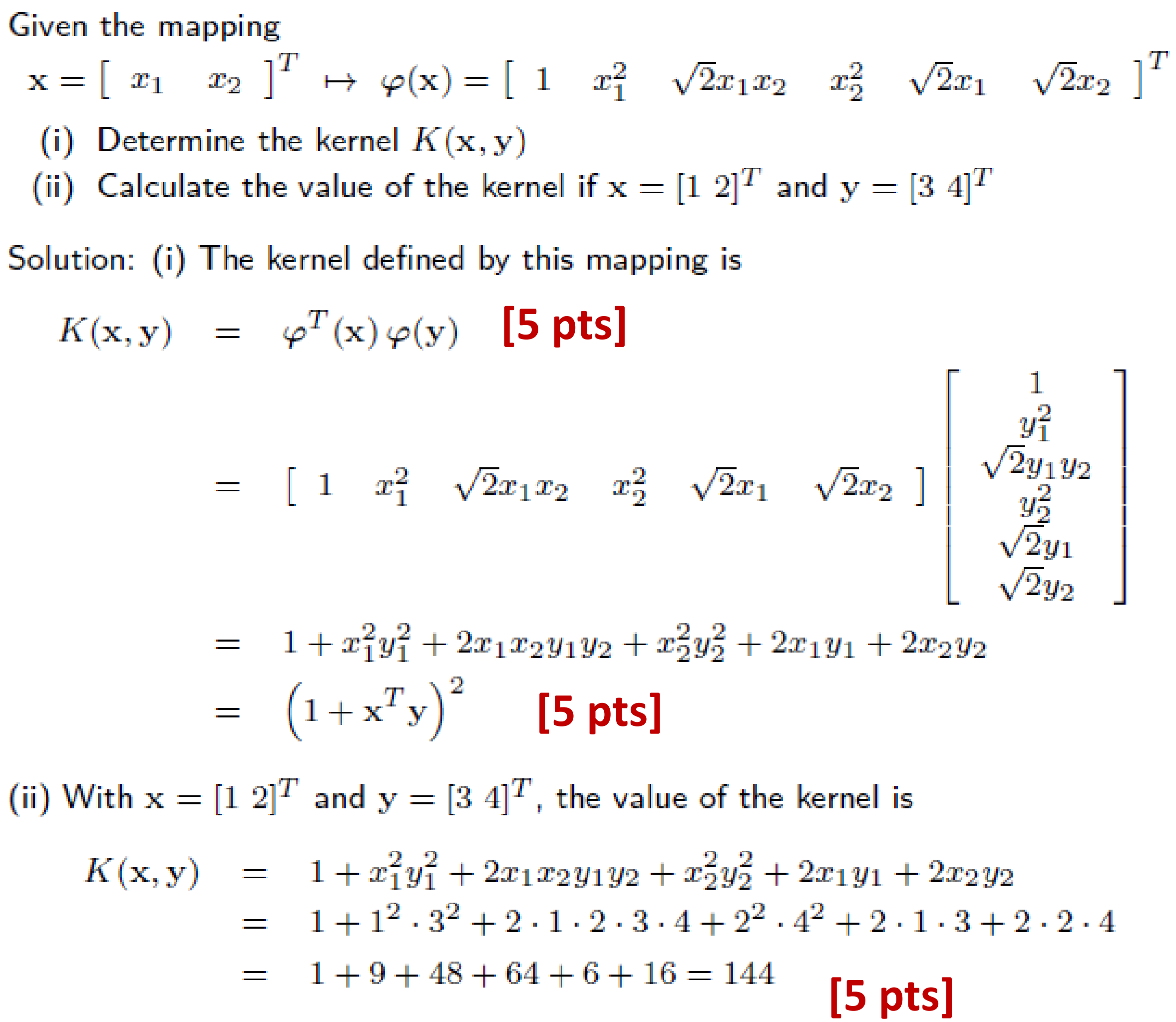
Berrak Sisman

Assistant Professor, ISTD Pillar, SUTD

Graded by Perry Lam

Support Vector Machines

Question 1. 1 [15 pts]



Given the mapping

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T \mapsto \varphi(\mathbf{x}) = \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix}^T$$

(i) Determine the kernel $K(\mathbf{x}, \mathbf{y})$

(ii) Calculate the value of the kernel if $\mathbf{x} = [1\ 2]^T$ and $\mathbf{y} = [3\ 4]^T$

Solution: (i) The kernel defined by this mapping is

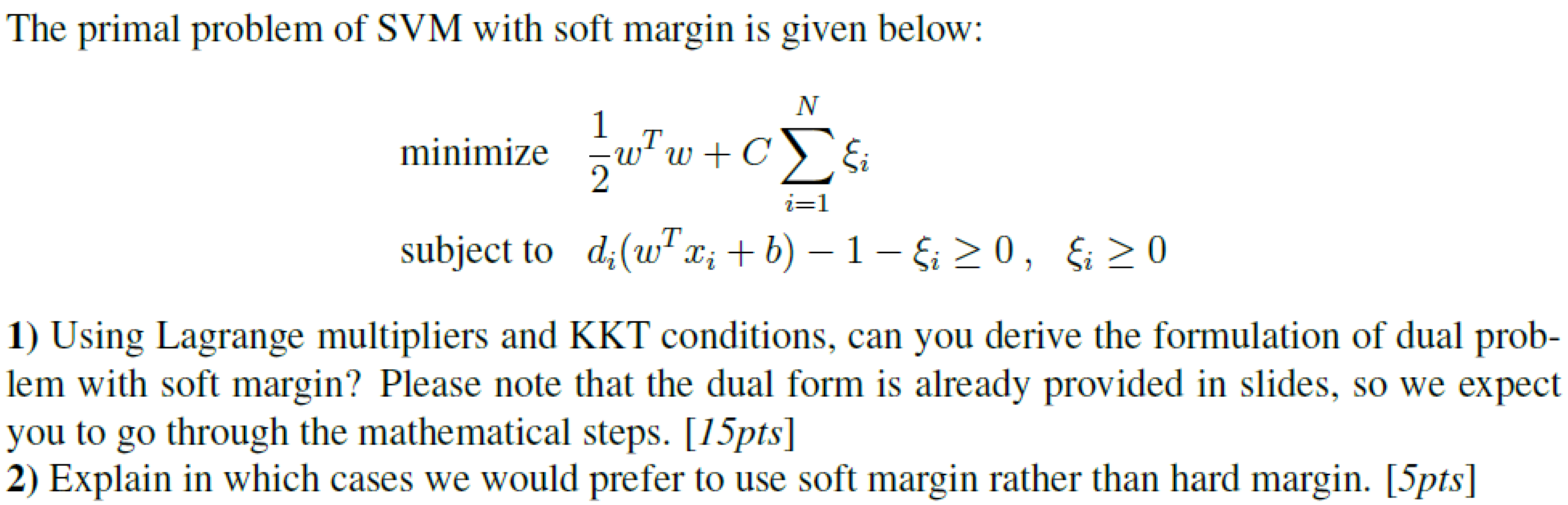
$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \varphi^T(\mathbf{x})\,\varphi(\mathbf{y}) \quad \textbf{[5 pts]} \\ &= \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \end{bmatrix} \\ &= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\ &= \left(1 + \mathbf{x}^T\mathbf{y}\right)^2 \quad \textbf{[5 pts]} \end{aligned}$$

(ii) With $\mathbf{x} = [1\ 2]^T$ and $\mathbf{y} = [3\ 4]^T$, the value of the kernel is

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\ &= 1 + 1^2 \cdot 3^2 + 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4 + 2^2 \cdot 4^2 + 2 \cdot 1 \cdot 3 + 2 \cdot 2 \cdot 4 \\ &= 1 + 9 + 48 + 64 + 6 + 16 = 144 \end{aligned}$$

[5 pts]

Question 1. 2 [20 pts]



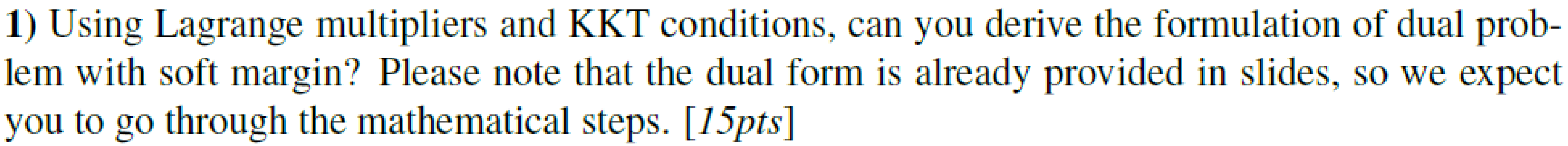
The primal problem of SVM with soft margin is given below:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \\ &\text{subject to} \quad d_i(w^T x_i + b) - 1 - \xi_i \geq 0\,, \quad \xi_i \geq 0 \end{aligned}$$

1) Using Lagrange multipliers and KKT conditions, can you derive the formulation of dual problem with soft margin? Please note that the dual form is already provided in slides, so we expect you to go through the mathematical steps. [*15pts*]
2) Explain in which cases we would prefer to use soft margin rather than hard margin. [*5pts*]

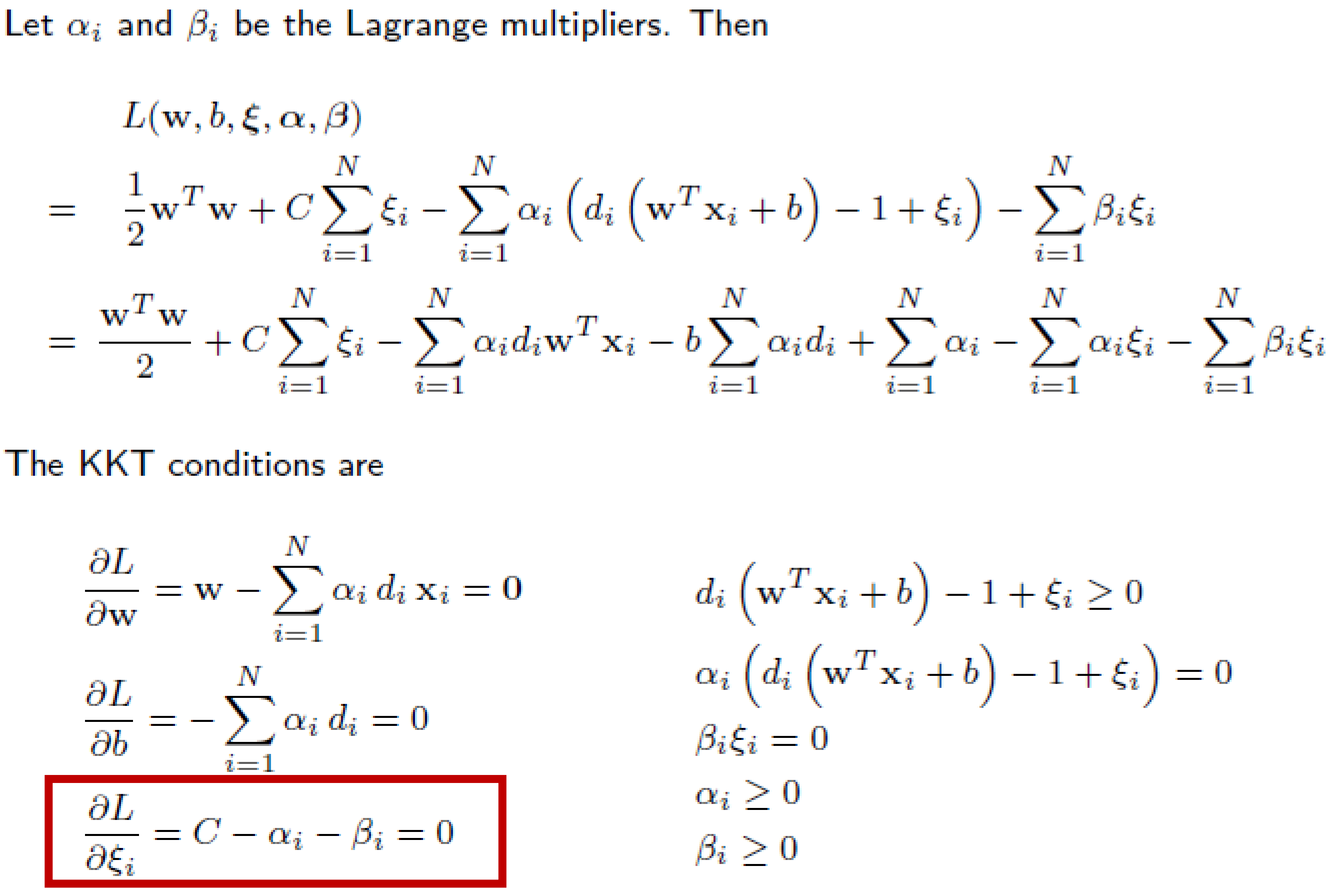
Question 1. 2 [20 pts]

[8 pts] for writing all KKT conditions;



1) Using Lagrange multipliers and KKT conditions, can you derive the formulation of dual problem with soft margin? Please note that the dual form is already provided in slides, so we expect you to go through the mathematical steps. [*15pts*]

Solution



Let α_i and β_i be the Lagrange multipliers. Then

$$\begin{aligned} & L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ = \; & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{N}\beta_i\xi_i \\ = \; & \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i d_i \mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N}\alpha_i d_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\beta_i\xi_i \end{aligned}$$

The KKT conditions are

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i\, d_i\, \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\alpha_i\, d_i = 0$$

Very important!

$$\boxed{\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0}$$

$$d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i \geq 0$$

$$\alpha_i\left(d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right) = 0$$

$$\beta_i\xi_i = 0$$

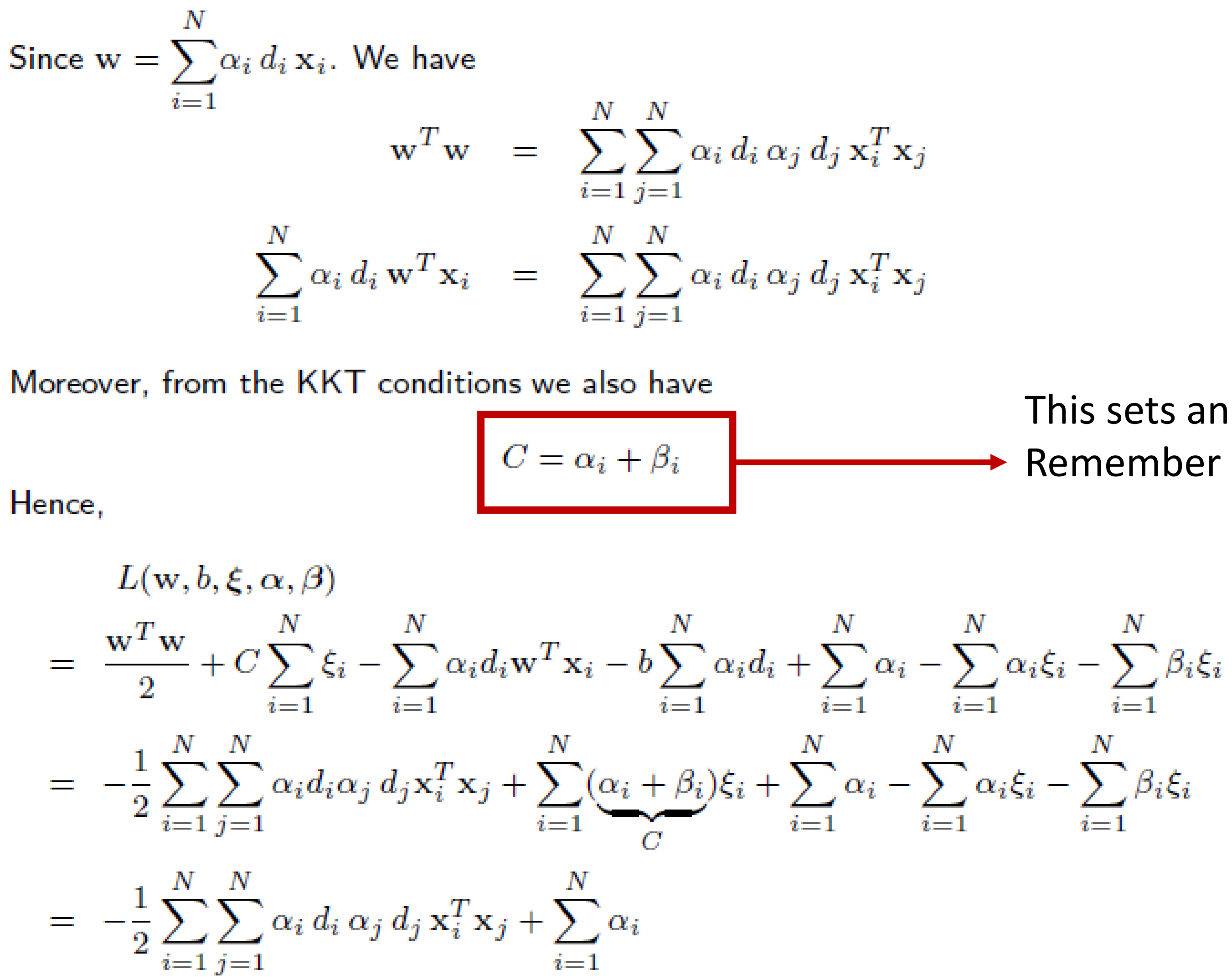
$$\alpha_i \geq 0$$

$$\beta_i \geq 0$$

Next slide

Question 1. 2 [20 pts]

Solution



Since $\mathrm{w} = \sum_{i=1}^{N} \alpha_i \, d_i \, \mathrm{x}_i$. We have

$$\begin{aligned} \mathrm{w}^T \mathrm{w} &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, d_i \, \alpha_j \, d_j \, \mathrm{x}_i^T \mathrm{x}_j \\ \sum_{i=1}^{N} \alpha_i \, d_i \, \mathrm{w}^T \mathrm{x}_i &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, d_i \, \alpha_j \, d_j \, \mathrm{x}_i^T \mathrm{x}_j \end{aligned}$$

Moreover, from the KKT conditions we also have

$$\boxed{C = \alpha_i + \beta_i}$$

Hence,

$$\begin{aligned} & L(\mathrm{w}, b, \xi, \alpha, \beta) \\ = \ & \frac{\mathrm{w}^T \mathrm{w}}{2} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i d_i \mathrm{w}^T \mathrm{x}_i - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i \\ = \ & -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j \, d_j \mathrm{x}_i^T \mathrm{x}_j + \sum_{i=1}^{N} \underbrace{(\alpha_i + \beta_i)}_{C} \xi_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i \\ = \ & -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, d_i \, \alpha_j \, d_j \, \mathrm{x}_i^T \mathrm{x}_j + \sum_{i=1}^{N} \alpha_i \end{aligned}$$

This sets an upper bound for α_i!
Remember that $\beta_i \geq 0$, and $\alpha_i = C - \beta_i$
$\mathbf{0 \leq \boldsymbol{\alpha_i} \leq C}$

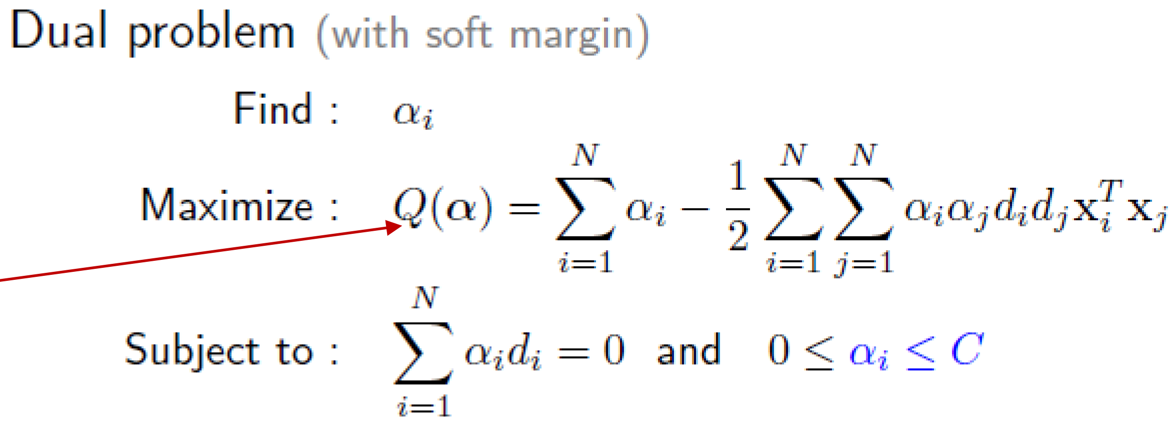
[7 pts] for solving the KKT conditions and obtaining the dual form;

Next slide

Question 1. 2 [20 pts]

Solution

Please note that $Q(\alpha)$ is same as that for the dual problem without soft margin.



Dual problem (with soft margin)

Find : α_i

Maximize : $Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$

Subject to : $\sum_{i=1}^{N} \alpha_i d_i = 0$ and $0 \leq \alpha_i \leq C$

If the previous 2 pages are correct, then the student gets 15 points.

Question 1. 2 [20 pts]

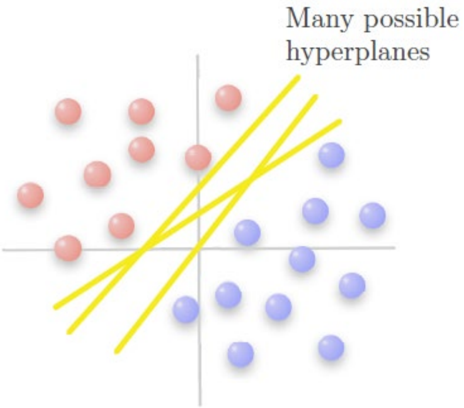
2) Explain in which cases we would prefer to use soft margin rather than hard margin. [5*pts*]

Solution

If our data is linearly separable, we can use hard margin. As discussed in lectures, hard margin is successful to handle such data. An example from our lecture notes is given below:

linearly separable

(Students do not need to include this)



[5 pts] students need to mention "not linearly separable" case

If our data is **not linearly separable**, we can use soft margin to allow a few points to be on the wrong side.

Question 1.3: Hands-on [15 pts]

Answer:

Kernel 0 (linear) accuracy = 74.6032% (47/63)

Kernel 1 (polynomial) accuracy = 53.9683% (34/63)

Kernel 2 (RBF) accuracy = 84.127% (53/63)

Kernel 3 (Sigmoid) accuracy = 79.3651% (50/63)

[12 pts] for correct accuracy

Question 1.3: Hands-on [15 pts]

Answer:

[3 pts] for choosing RBF

Kernel 0 (linear) accuracy = 74.6032% (47/63)

Kernel 1 (polynomial) accuracy = 53.9683% (34/63)

Kernel 2 (RBF) accuracy = 84.127% (53/63) **Best performance!**

Kernel 3 (Sigmoid) accuracy = 79.3651% (50/63)

Logistic Regression

Question 2.1 [20 pts]

Suppose that you have trained a logistic regression classifier, and it outputs on a new example a prediction $h_\theta(x) = 0.35$. This means (check all that apply):

Solution

1) Our estimate for $P(y = 0|x; \theta)$ is 0.35 **NO** **[5 pts]**

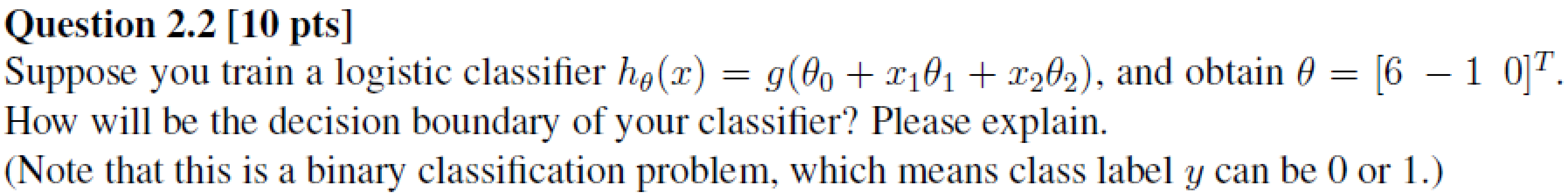
2) Our estimate for $P(y = 0|x; \theta)$ is 0.65 **YES** **[5 pts]**

3) Our estimate for $P(y = 1|x; \theta)$ is 0.35 **YES** **[5 pts]**

4) Our estimate for $P(y = 1|x; \theta)$ is 0.65 **NO** **[5 pts]**

- **Note that $h_\theta(x)$ is the estimated probability that $y = 1$ on input x. If $h_\theta(x)$=0.35, it means that $P(y = 1|x; \theta) = 0.35$.**
- **Please also note that $p(y = 1|x; \theta)$+$p(y = 0|x; \theta) = 1$. So $p(y = 0|x; \theta)$ will be equal to 0.65.**

Question 2.2 [10 pts]



Question 2.2 [10 pts]
Suppose you train a logistic classifier $h_\theta(x) = g(\theta_0 + x_1\theta_1 + x_2\theta_2)$, and obtain $\theta = [6 \;\; -1 \;\; 0]^T$.
How will be the decision boundary of your classifier? Please explain.
(Note that this is a binary classification problem, which means class label y can be 0 or 1.)

Solution

Let’s try to figure out where the hypothesis ends of predicting $y = 0$ and $y = 1$

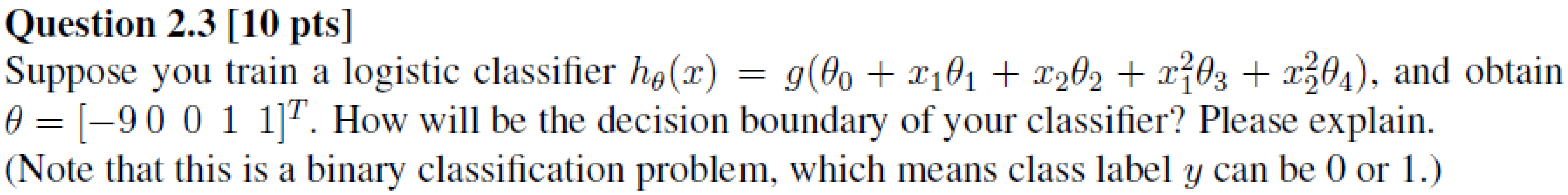
$y = 1$ if $6 - x_1 \geq 0$ $\qquad\qquad$ $y = 0$ if $6 - x_1 < 0$

$$\boxed{\boldsymbol{y = 1 \text{ if } 6 \geq x_1 \qquad\qquad y = 0 \text{ if } 6 < x_1}}$$

[10 pts]

(Please note that decision boundary is a property of hypothesis and the parameters…)

Question 2.3 [10pts]



Question 2.3 [10 pts]
Suppose you train a logistic classifier $h_\theta(x) = g(\theta_0 + x_1\theta_1 + x_2\theta_2 + x_1^2\theta_3 + x_2^2\theta_4)$, and obtain $\theta = [-9\,0\;\;0\;\;1\;\;1]^T$. How will be the decision boundary of your classifier? Please explain.
(Note that this is a binary classification problem, which means class label y can be 0 or 1.)

Solution

Let's try to figure out where the hypothesis ends of predicting $y = 0$ and $y = 1$

$y = 1$ if $\; -9 + x_1^2 + x_2^2 \geq 0$ $\qquad$ $y = 0$ if $\; -9 + x_1^2 + x_2^2 < 0$

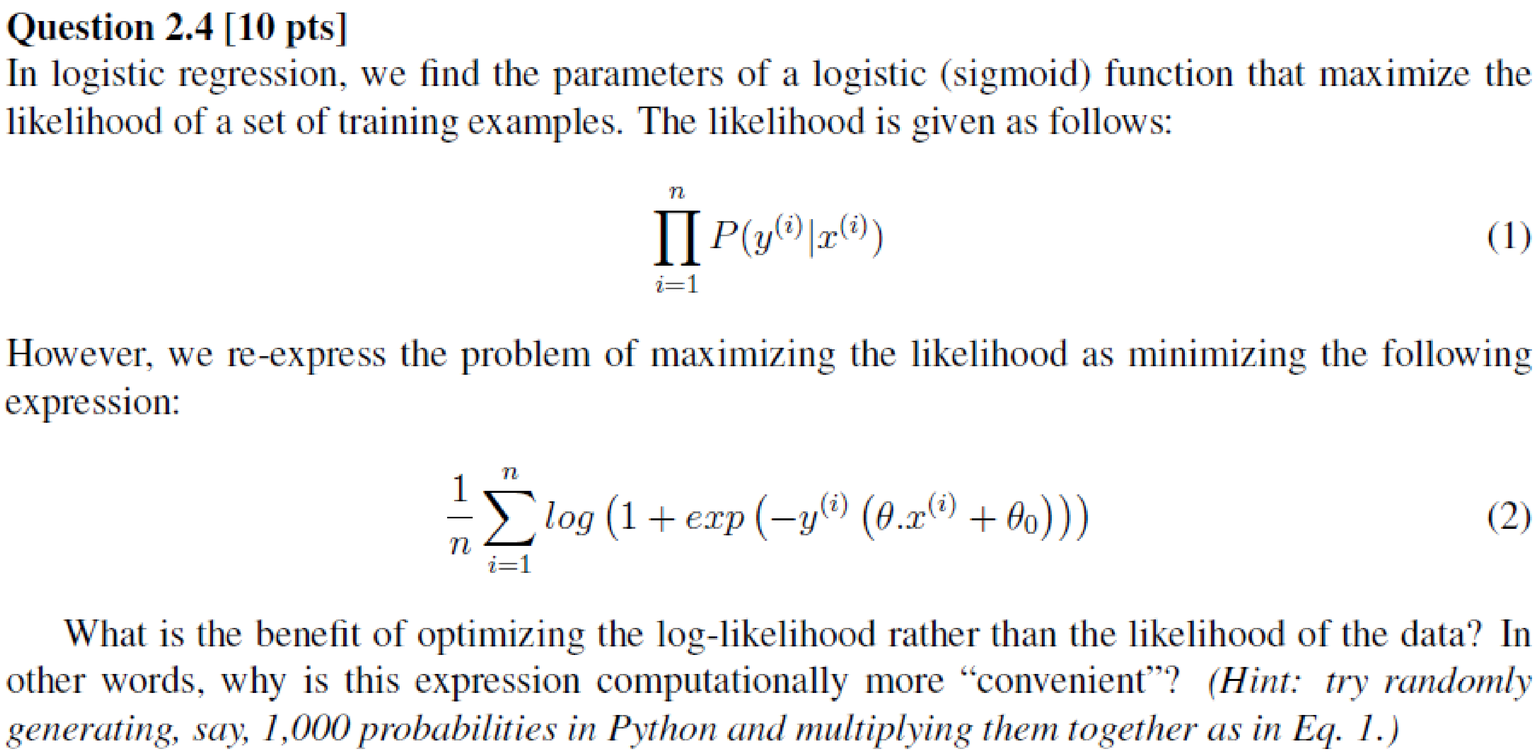
$$\boxed{\boldsymbol{y = 1 \text{ if } \quad x_1^2 + x_2^2 \geq 9 \qquad\qquad y = 0 \text{ if } \quad x_1^2{+}x_2^2 < 9}}$$

[10 pts]

(Please note that decision boundary is a property of hypothesis and the parameters…)

How does the decision boundary look like? It is a circle with r=3! **[not compulsory]**

Question 2.4 [10 pts]



Question 2.4 [10 pts]

In logistic regression, we find the parameters of a logistic (sigmoid) function that maximize the likelihood of a set of training examples. The likelihood is given as follows:

$$\prod_{i=1}^{n} P(y^{(i)}|x^{(i)}) \tag{1}$$

However, we re-express the problem of maximizing the likelihood as minimizing the following expression:

$$\frac{1}{n}\sum_{i=1}^{n} log\left(1 + exp\left(-y^{(i)}\left(\theta.x^{(i)} + \theta_0\right)\right)\right) \tag{2}$$

What is the benefit of optimizing the log-likelihood rather than the likelihood of the data? In other words, why is this expression computationally more "convenient"? *(Hint: try randomly generating, say, 1,000 probabilities in Python and multiplying them together as in Eq. 1.)*

Answer: [10 pts]

Progressively multiplying many probabilities together as in Equation 1 quickly gives a result that is too small to be representable in computer memory (this is known as an underflow problem). In contrast, Equation 2 uses a sum over terms that makes this problem less likely to occur.