

# Alan Copa

## Machine Learning Assignment 1 Linear Models & Kernel Methods

Submission deadline: October 26, 2024

### 1 Problem 1. Ridge Regression (10 points)

In a regression task, we have vectors  $\mathbf{x} \in \mathbb{R}^D$ , target values  $y \in \mathbb{R}$  associated with them, and some model  $f(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$  to predict the target values for arbitrary vectors in  $\mathbb{R}^D$ .

Suppose we have a training dataset  $\{\Phi, \mathbf{t}\}$ , where  $\Phi \in \mathbb{R}^{N \times D}$  is the design matrix in which each row is a feature vector  $\phi(\mathbf{x})$  of a training point  $\mathbf{x}$ , and  $\mathbf{t} \in \mathbb{R}^{N \times 1}$  is the vector with target values for the training points.  $N$  is the number of points in the training dataset, and  $D$  is the dimensionality of the feature space. Suppose that each entry in the last column of  $\Phi$  is equal to 1. Your task is to derive the closed form solution for the optimal parameters of a ridge regression model.

- State the equation of a ridge regression model and identify the model parameters
- State the equation for the loss function (mean squared error) with an  $\ell_2$  regularization weighted by  $\lambda$
- State which condition should be met in order to find the model parameters
- Find the ideal model parameters under the proposed loss function.

*Note:* You can use  $\|\cdot\|$  as the Euclidean norm of a vector;  $\phi(\mathbf{x}_n)$  and  $t_n$  are the  $n$ -th rows of  $\Phi$  and  $\mathbf{t}$  respectively.

### 2 Problem 2. Feature Engineering (10 points) and Basic Concepts (10 points)

Suppose you have the following set  $S$  of 2D points,  $S_n = (x_n^{(1)}, x_n^{(2)})$ .

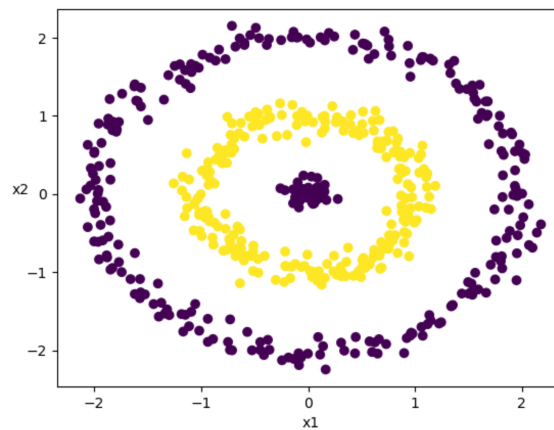


Figure 1: Color denotes the class attribution of a point: blue points belong to the class  $C_1$ , yellow points belong to the class  $C_2$ .

- Explain in detail 2 classification algorithms that could solve the problem. Discuss advantages and disadvantages of each.
- Propose new features for points in  $S$  based on  $x^{(1)}$  and  $x^{(2)}$ . In this new feature space, classes  $C_1$  and  $C_2$  should be linearly separable. Come up with 2 different solutions meeting the stated criteria.

### Problem 3. Kernel Functions (10 points)

Consider the following function  $f : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ :

$$f(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{x})(\mathbf{x}^T \mathbf{y})(\mathbf{y}^T \mathbf{y})$$

Prove that  $f$  is a valid kernel or prove the opposite.

The only rules allowed to use without proof are the following:

- Kernel functions are *linear* and positive.
- Kernel functions can be expressed as an inner product
- A kernel function of 2 inputs can be expressed as another kernel of a transformation of those inputs (into a potentially different space).

Begin by formalizing those rules and apply them to prove or disprove the statement.

### Problem 4. SVM (10 points)

Consider the following training data:

Class	$x_1$	$x_2$
+	1	1
+	2	2
+	0	2
−	1	−1
−	−1	0
−	0	0

1. Plot the six training points. Are the classes  $\{+, -\}$  linearly separable?
2. Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors.
3. If you remove one of the support vectors, does the size of the optimal margin decrease, stay the same, or increase?
4. Is your answer to (3) also true for any dataset? Provide a counterexample or give a short proof.

# 1 Problem 1. Ridge Regression (10 points)

In a regression task, we have vectors  $\mathbf{x} \in \mathbb{R}^D$ , target values  $y \in \mathbb{R}$  associated with them, and some model  $f(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}$  to predict the target values for arbitrary vectors in  $\mathbb{R}^D$ .

Suppose we have a training dataset  $\{\Phi, \mathbf{t}\}$ , where  $\Phi \in \mathbb{R}^{N \times D}$  is the design matrix in which each row is a feature vector  $\phi(\mathbf{x})$  of a training point  $\mathbf{x}$ , and  $\mathbf{t} \in \mathbb{R}^{N \times 1}$  is the vector with target values for the training points.  $N$  is the number of points in the training dataset, and  $D$  is the dimensionality of the feature space. Suppose that each entry in the last column of  $\Phi$  is equal to 1. Your task is to derive the closed form solution for the optimal parameters of a ridge regression model.

- State the equation of a ridge regression model and identify the model parameters
- State the equation for the loss function (mean squared error) with an  $\ell_2$  regularization weighted by  $\lambda$
- State which condition should be met in order to find the model parameters
- Find the ideal model parameters under the proposed loss function.

*Note:* You can use  $\|\cdot\|$  as the Euclidean norm of a vector;  $\phi(\mathbf{x}_n)$  and  $t_n$  are the  $n$ -th rows of  $\Phi$  and  $\mathbf{t}$  respectively.

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_D(x_1) & 1 \\ \vdots & \vdots & & & 1 \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_D(x_N) & 1 \end{pmatrix} \quad \begin{matrix} N \times 1 \\ \mathbf{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{matrix}$$

$$X^{D \times 1} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$$

$$W = \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix}$$

a) Ridge regression model

$\hat{y}$ : predicted target values

$\Phi$ : design matrix

$w$ : model parameters

$$\hat{y} = \Phi w$$

b)  $E(w) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2} (w^T \phi(x_n) - t_n)^2$  MSE

MSE loss function with  $\ell_2$  reg. term.:

$$\mathcal{L}(w) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2} (w^T \phi(x_n) - t_n)^2 + \frac{\lambda}{2} \|w\|_2^2$$

c) Condition to find the model parameters:

$$\frac{\partial L(w)}{\partial w} = 0$$

d) optimal parameters

$$L(w) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2} (w^T \phi(x_n) - t_n)^2 + \frac{\lambda}{2} \|w\|_2^2$$

$$\frac{\partial L}{\partial w} = \frac{1}{N} \sum_{n=1}^N \left( \frac{1}{2} \cdot 2 (w^T \phi(x_n) - t_n) \cdot \phi_j(x_n) \right) + \lambda w$$

$$= \frac{1}{N} \sum_{n=1}^N (w^T \phi(x_n) - t_n) \phi_j(x_n) + \lambda w$$

$$= \frac{1}{N} \left( w^T \sum_{n=1}^N \phi(x_n) \phi^T(x_n) - \sum_{n=1}^N t_n \phi^T(x_n) \right) + \lambda w$$

$$= \frac{1}{N} (w^T \Phi^T \Phi - T^T \Phi) + \lambda w \stackrel{!}{=} 0$$

$$w^T \Phi^T \Phi + \lambda w = \tau^T \Phi$$

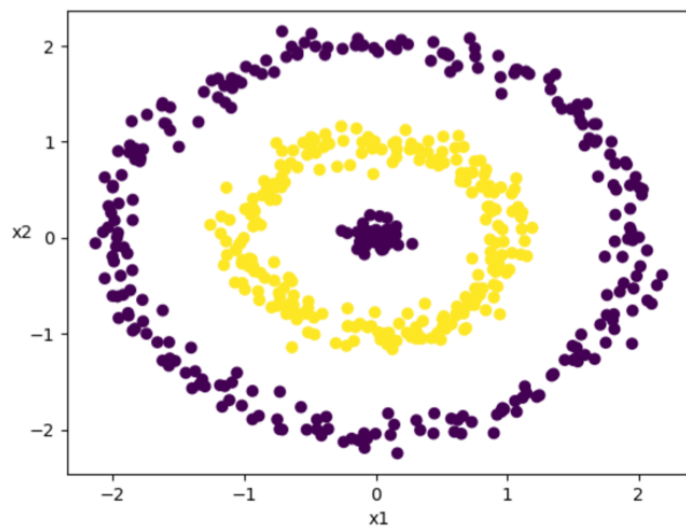
$$\Phi^T \Phi w + \lambda w = \Phi^T \tau$$

$$(\Phi^T \Phi + \lambda I) w = \Phi^T \tau$$

$$w_{\min} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \tau$$

## 2 Problem 2. Feature Engineering (10 points) and Basic Concepts (10 points)

Suppose you have the following set  $S$  of 2D points,  $S_n = (x_n^{(1)}, x_n^{(2)})$ .



$$z = x_1^2 + x_2^2$$

Figure 1: Color denotes the class attribution of a point: blue points belong to the class  $C_1$ , yellow points belong to the class  $C_2$ .

- a • Explain in detail 2 classification algorithms that could solve the problem. Discuss advantages and disadvantages of each.
- b • Propose new features for points in  $S$  based on  $x^{(1)}$  and  $x^{(2)}$ . In this new feature space, classes  $C_1$  and  $C_2$  should be linearly separable. Come up with 2 different solutions meeting the stated criteria.

## a) non parametric kNN classifier:

This algorithm consider the closest  $k$  data points to determine by majority to which class the current data point has to be assigned.

kNN steps:

- ① set the number of  $k$  neighbors
- ② calculate distance of current point from all other points of the dataset
- ③ find the  $k$  nearest neighbors of the point
- ④ identify the class of the point by looking at the majority of the  $k$  nearest neighbors

advantages:

- no need to train the model
- non parametric, can easily adapt to non linearity
- no assumptions about data distribution

disadvantages:

- need to compute all distances / difficult to define a metric for closeness
- sensitive to outliers
- does not discover structure
- problematic in high dimensional space

# Logistic regression

Logistic regression is a linear model used for binary classification. It works by modeling the relationship between the input features and the probability of a point to belong to a specific class.

This model uses a sigmoid function to convert the linear output into a probability (from 0 to 1). During training the algorithm adjusts the weights  $w$  and the bias  $b$  to minimize the log loss function, which penalizes the incorrect predictions.

$$P(y=1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

## advantages:

- training quick and less computing power needed
- probabilistic output useful for confidence level

## disadvantages



- not suitable for this case:  
the classes are not linearly separable  
unless we introduce non linearity  
with the features
- Sensitive to outliers
- no closed form to compute  $w$  and  $b$

b) "Circle" feature introduction

$$z = \sqrt{(x^{(1)})^2 + (x^{(2)})^2}$$

this will introduce a circle/ring based feature  
and open to non linearity in the model.

The decision boundary will follow a circle  
shape classifying better the points.

New polynomial feature

$$z_1 = (x^{(1)})^2 + (x^{(2)})^2 \quad z_2 = x^{(1)} \cdot x^{(2)}$$

$z_1$  introduces the squared radial distance  
from the origin that combined with  $z_2$   
can imitate the ring behaviour of the  
distribution of the data

### Problem 3. Kernel Functions (10 points)

$$g(x) = (\|x\|^2, x)$$

Consider the following function  $f: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ :

$$f(x, y) = (x^T x)(x^T y)(y^T y)$$

Prove that  $f$  is a valid kernel or prove the opposite.

The only rules allowed to use without proof are the following:

- a) • Kernel functions are *linear* and positive.
- b) • Kernel functions can be expressed as an inner product
- c) • A kernel function of 2 inputs can be expressed as another kernel of a transformation of those inputs (into a potentially different space).

Begin by formalizing those rules and apply them to prove or disprove the statement.

a)  $k(x, y) = c k_1(x, y)$   
 $k(x, y) = k_1(x, y) + k_2(x, y)$

$c \geq 0$

b)  $k(x, y) = \phi(x)^T \phi(y)$

$\phi(x), \phi(y)$  transformed  
feature vectors

c)  $k'(x, y) = k(g(x), g(y))$

$g$  transformation  
function to a  
new space

To prove that  $f$  is a valid kernel we need to show that  $f$  is symmetric, positive semi-definite and that it can be expressed as an inner product.

① Symmetric

$$f(x, y) \stackrel{?}{=} f(y, x)$$

$$(x^T x) (x^T y) (y^T y) \stackrel{?}{=} (y^T y) (y^T x) (x^T x)$$

yes, since  $y^T x = x^T y$

② positive semi-definite

for any vector  $c \in \mathbb{R}$   $c^T K c \geq 0$   
must be satisfied.  $K$  is the kernel  
matrix

$$\text{let's take } x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$f(x_1, x_2) = (1+0)(0+0)(0+1) = 0 = f(x_2, x_1)$$

$$f(x_2, x_2) = 1 \cdot 1 \cdot 1 = 1$$

$$f(x_1, x_1) = 1 \cdot 1 \cdot 1 = 1$$

$$K = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$c = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$c^T K$$

$$c \stackrel{?}{\geq} 0$$

$$\begin{aligned}
 & [1 \ -2] \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \\
 & = [-2 \ 1] \begin{bmatrix} 1 \\ -2 \end{bmatrix} = -2 - 2 = \textcircled{-4} \quad \angle 0 \quad \nabla
 \end{aligned}$$

$\Rightarrow f$  is not a valid kernel.

③ express  $f$  as a inner product:

there is no way to define  
 $f$  as a inner product  $\phi(x)^T \phi(y)$

$$f(x, y) = \|x\|^2 \phi(x)^T \phi(y) \|y\|^2$$

$\|x\|^2$  and  $\|y\|^2$  introduce non linearity.

Furthermore:

- it doesn't exist any linear combination of valid kernel able to create  $f$  since it is a non linear combination of inner products.

• we could express  $f$   
with  $g(x) = (\|x\|^2, x)$  s.t.

$$f(x, y) = k'(g(x), g(y))$$

but  $k'$  with  $g$  doesn't  
preserve its linearity