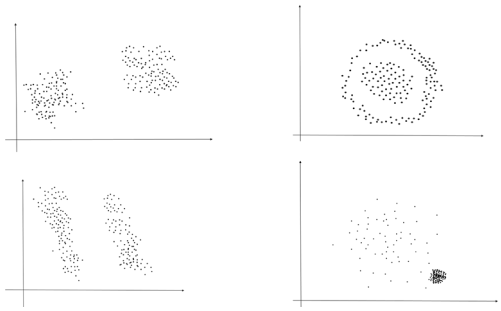
Alireza Gargoori Motlagh, Sina Mazaheri, Hadis Ahmadian, Taha Akbari

1 Clustering and K-means (35 points)

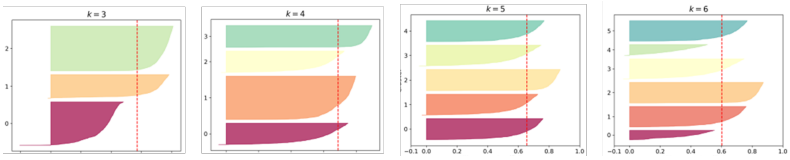
1.1 Intuitive Understanding

- In each sample below, draw the boundry that K-means finds for $K = 2$. Do you think the clusters separated by borders found by K means is meaningfull in each case? If not, what property of data causes this?



- In which case above, feature scaling can solve the problem? Why?

1.2 Finding proper value of K



- For finding a proper value for K, what is the advantage of using silhouette score in comparison to inertia?
- In the figure above, you can see silhouette diagram for different values of k. which value for K is better? Why?

1.3 Applications of Clustering

Two applications of clustering are Active learning and semi-supervised learning. Explain usage of clusting in each of these approaches and the diffrences between them briefly.

2 Whitening Using PCA (25 points)

Whitening is one of the pre-processing techniques which is used in practical ML. By whitening, we mean that for a given feature matrix $\mathbf{X}$, this feature matrix should have zero mean vector and identity matrix as its covariance matrix. Explain that how we could transform $\mathbf{X}$ by using its covariance matrix principal components in order to have a whitened data set? After that, prove this new feature matrix has the desired properties, namely, zero mean vector and identity matrix as its covariance matrix.

3 Linear Regression

3.1 Lagrange Multipliers (25 points)

Using Lagrange Multipliers give a geometric interpretation of why Lasso Regression results in more sparse solutions. Consider using $L_p$where $p < 1$ instead of $p = 1$. Show that we will get even more sparse solutions when we decrease p. Additionally explain why L_1 norm is used instead of smaller values for getting sparse solutions. (Hint: Analysis the convexity of unit ball with L_p norm.)

3.2 Ridge Regression (optional) (10 points)

Prove that the optimal value for the Ridge regression is $\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ and can be obtained by minimizing the Error function below:

$$L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \frac{1}{2}\lambda\|\mathbf{w}\|_2^2 \tag{1}$$

4 Generalization Error (15 points)

- What is the effect of dimentionality reduction on overfitting? Why?
- We know that one solution for overfitting is using more data for training. However finding more data is not always possible, in this case we can alter the data that we already have so that by changing them slightly, we generate new data. This is called data augmentation. Search about data augmentation methods and give three examples of them.

5 Kernels

5.1 Feature Space (15 points)

Consider $x, y \in \mathbb{R}^2$, the kernel function defined as

$$k(x, y) = (1 + cos(\angle x, y))^2,$$

where $cos(\angle x, y)$ means the cosine of the angle between x and y vectors. Find the feature space $\phi(x)$ which corresponds to this kernel.

5.2 Kernel Matrix (20 points)

5.2.1

Suppose for now that k is indeed a valid kernel corresponding to some feature mapping ϕ. Now, consider some infinite set of m points, $\{x^{(1)}, ..., x^{(m)}\}$, and let a square m-by-m matrix K be defined so that its (i, j)-entry is given by $K_{i,j} = k(x^{(i)}; x^{(j)})$. This matrix is called the Kernel matrix.
Prove that K is a positive semi-definite matrix.

5.2.2

The positive semi-definitivity of kernel matrix turns out to be not only a necessary, but also a sufficient, condition for k to be a valid kernel, also called a Mercer kernel.
Prove that if K(kernel matrix) is a positive semi-definite matrix, then k is a dot product: $\exists\phi$ such that $k(x, y) = \phi(x).\phi(y)$
(Hint: Use diagonalization for real symmetric matrices.)

6 Feature Expansion (20 points)

6.1

For each dataset below in figure 1, find a 1-dimensional transform $\phi : R^d \rightarrow R$ such that in the new feature space, 2 classes are linearly separable.

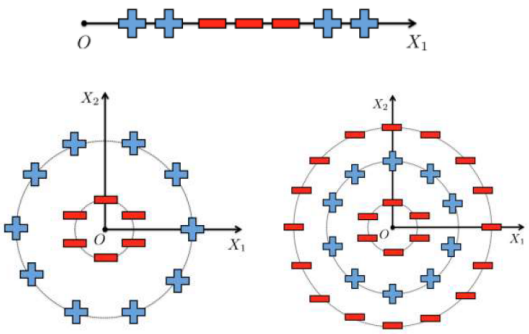


Figure 1: Datasets in original feature space

6.2

As you know, the equation of a circle in 2-d plane is defined as $(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$. By expanding the former equation, show that any circular region is linearly separable in (x_1, x_2, x_1^2, x_2^2) feature space.

7 SVM Decision Boundaries (30 points)

In Figure 2 we have plotted the decision boundaries and margins for SVM learned on the same data set using the following parameters (not in the same order as the figures):

(i) Linear kernel, C = 0.1

(ii) Linear kernel, C = 1

(iii) Linear kernel, C = 10

(iv) RBF kernel, γ = 0.1, C = 15

(v) RBF kernel, γ = 1, C = 3

(vi) RBF kernel, γ = 10, C = 1

Match each one of the figures with one of these parameter settings. Explain your matchings in few sentences.

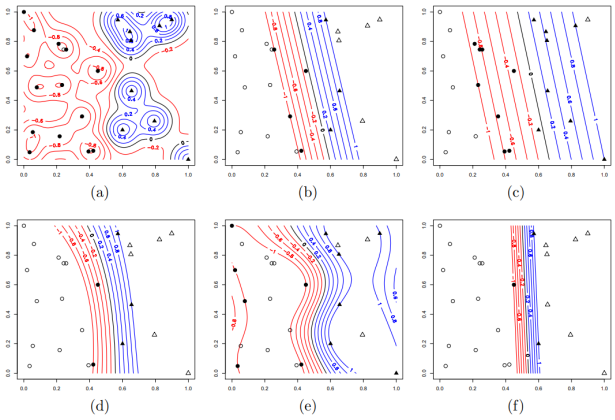


Figure 2: Circles and triangles denote Class 1 and 2 respectively, solid points are support vectors.

8 Monte Carlo Cross Validation (15 points)

In Monte Carlo Cross Validation(MCCV), also known as repeated random subsampling CV, you randomly select (without replacement) some fraction of your data to form the training set, and then assign the rest of the points to the validation set. This process is then repeated multiple times, generating (at random) new training and validation partitions each time. In other words, it simply splits the N data points into the two subsets n_t and n_v by sampling, without replacement. There exists $\binom{N}{n_t}$ unique training sets, but MCCV avoids the need to run this many iterations.

8.1

How does the choice of n_t affects the bias-variance trade-off?

8.2

Compare MCVV with k-Fold Cross Validation and state its pros and cons.