6 Kernel Methods

Exercise 6.1

Comment the following statements about adding new features to your model:

1. It is always a good idea to add some feature in classification since they increase the chance to consider feature spaces where it is possible to linearly separate the classes;

2. The addition of new features requires a longer time for the training of the model;

3. The addition of new features requires a longer time in prediction of newly seen samples;

4. It is not a trivial task to chose properly the features which might improve your learner capabilities;

5. You need to know the right set of features if we want to make use of them.

Exercise 6.2

Answer the following questions about kernels. Motivate your answers.

1. Can you define a kernel over a feature set composed of colors? For instance the set could be $\mathcal{F} = \{red, green, blue, black, white\}$.

2. Can you define a kernel over a feature set composed of graphs?

3. Do you prefer to have a larger hard drive and/or a faster CPU to apply a kernel method?

4. Assume to have a non-linearly separable dataset, but you know which mapping is able project them in a linearly separable space. Are there still reasons to consider the use of kernels?

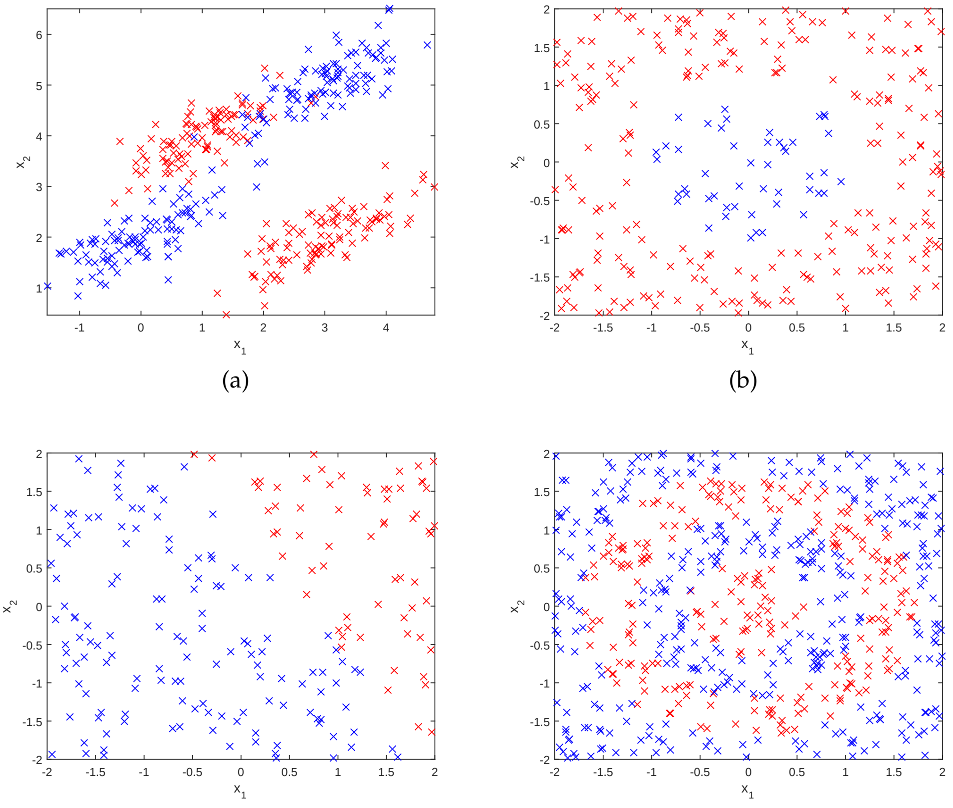
* Exercise 6.3

Derive the kernel formulation for the ridge regression, when we consider $\phi(x)$ as input features.

Is $k(x, x') = \phi(x)^T \phi(x') + \lambda I$ always a valid kernel?.

Exercise 6.4

For which one of the dataset in Figure 6.1 you would use the kernel trick to represent your data? Would you use some other methodology? Provide motivation for your choice.



(a) (b)

(c) (d)

Figure 6.1: Different datasets.

Exercise 6.5

Consider $x, y \in \mathbb{R}^d$, which ones of these are similarity measure:

1. $k(x, y) = x^T y$ (dot product);
2. $k(x, y) = x^T y + (x^T y)^2$;
3. $k(x, y) = ck_1(x, y) + k_2(x, y) \times k_3(x, y)$, where k_1, k_2 and k_3 are valid kernels in $\mathbb{R}^d$;

4. $k(x, y) = \log(x)e^{-y}$ ($d = 1$);
5. $k(x, y) = x^T Ay$ with $A = \begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix}$ ($d = 2$);
6. $k(x, y) = \sqrt{(1 - \cos^2(x))} \cos(y - \pi/2)$, ($d = 1$).

Exercise 6.6

Tell if the following functions are valid kernels. Motivate your answers. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

1. $k_1(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} + \mathbf{x}^T\mathbf{1} + \mathbf{y}^T\mathbf{1} + d$, where $\mathbf{1} \in \mathbb{R}^d$ is the vector of all ones.
2. $k_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} - \|\mathbf{x}\|^2$
3. $k_3(\mathbf{x}, \mathbf{y}) = k_1(\cos(\mathbf{x}), \cos(\mathbf{y}))^3$, where the $\cos(\cdot)$ function is applied element-wise.
4. $k_4(\mathbf{x}, \mathbf{y}) = \exp\left(k_2(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{y}, \mathbf{x})\right)$

Exercise 6.7

Suppose you want to use a GP for a regression problem. You know that it varies a lot in some dimensions and less in others.

1. Which kind of covariance kernel would you use? Provide the analytic form of the kernel and motivate why you would choose it.
2. There exist other techniques which are able to handle this problem? Are there any drawbacks in doing so?
3. Why you should not consider such a model in the case you have the information that each dimension is equivalent to the others.

Exercise 6.8

Comment the following statements about GPs. Motivate your answers.

1. The more the samples we have in a point of the input space x the more it is likely that the variance of the process decreases in x.
2. We can choose any kind of prior distribution for a GP and we are assured to reach the true function if we get enough samples.
3. Gaussian process can be used only for regression problems.
4. Far from the region where we have points the variance of the GP gets larger and larger.

5. As in linear models, we are considering different variance for the noise in each point of the input space x.

Exercise 6.9

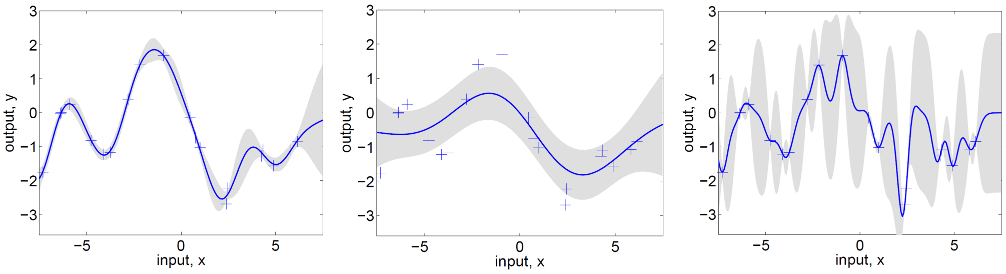
Associate the following set of parameters:

1. $\phi = 1$, $l = 1$ and $\sigma = 0.1$;

2. $\phi = 1.08$, $l = 0.3$ and $\sigma = 0.000005$;

3. $\phi = 1.16$, $l = 3$ and $\sigma = 0.89$;

of the Gaussian covariance $k(x, x') = \phi \, \exp\left(-\frac{1}{2l}(x - x')^2\right) + \sigma^2$ with the following figures:



(a)

(b)

(c)

Figure 6.2: Different GPs.

where the shaded areas represent the confidence intervals at 95%.

Provide motivations for your answers.

Exercise 6.10

Comment the following statements about GPs. Motivate your answers.

1. Differently from linear models, we are considering different variance for the noise in each point of the input space.

2. The specific GP formulation allows one to use them only for classification problems.

3. Any finite subset of the points in the output space predicted by a GP follows a Gaussian multivariate distribution.

4. If we have few samples in a portion of the input space it is likely that the GP will have high uncertainty in that region.

Exercise 6.11

Comment on the following statements about Gaussian Processes (GP). Assume to have a dataset generated from a GP $D = (x_i, y_i)_{i=1}^N$. Motivate your answers.

1. GPs are parametric methods.
2. The computation of the estimates of the variance of the GP $\hat{\sigma}^2(x)$ corresponding to the input x provided by D does not require the knowledge of the samples output $(y_1, \dots, y_N)$.
3. In the neighbourhood of the input points $(x_1, \dots, x_N)$, we observed the variance of the GP gets smaller and smaller as we collect more samples.
4. The complexity of the computation of the estimates of the mean $\hat{\mu}(x)$ and variance $\hat{\sigma}^2(x)$ scales as N^3, i.e., cubically with the number of samples N.

* Exercise 6.12

Assume to model a phenomenon $y : [0, 10] \to \mathbb{R}$ with a Gaussian Process. Assume to have a prior of $\text{GP}(0, k(x, x'))$ with $k(x, x') = \frac{1}{2} e^{-\frac{(x-x')^2}{2l}}$, lengthscale parameter $l = 2$, and a noise variance of $\sigma^2 = 2$. You have a single noisy sample from the real function $x_1 = 1$ and $y_1 = 5$.

1. Compute the value for the prior variance for the point $x_2 = 6$. Is it different from any other point in the input space?
2. Compute the value for the posterior mean and variance for the point $x_3 = 3$ given the sample (x_1, y_1).
3. Do you think that there exists another method able to provide a (meaningful) prediction for a regression problem only given a single point? Provide either a method or a class of methods able to do that.

Recall that the formulas to compute the posterior mean and variance of a GP are the following:

$$\mu(x) = k(x, x_t)(K_t + \sigma^2 I)^{-1} y_t,$$
$$s^2(x) = k(x, x) - k(x, x_t)^\top (K_t + \sigma^2 I)^{-1} k(x, x_t),$$

where K_t is the gram matrix (kernel built on the training data).

Exercise 6.13

Which of the following statements are true?

1. Suppose you have 2D input examples (i.e., $\mathbf{x_i} \in \mathbb{R}^2$). The decision boundary of the SVM (with the linear kernel) is a straight line.
2. If you are training multi-class SVM with the one-vs-all method, it is not possible to use a kernel.
3. The maximum value of the Gaussian kernel is 1.
4. If the data are linearly separable, an SVM using a linear kernel will return the same parameters $\mathbf{w}$ regardless of the chosen value of C.

Exercise 6.14

Are the following statements about Support Vector Machines (SVMs) True or False? Motivate your answers.

1. When using an SVM, the computational cost of computing predictions scales with the size of training samples.
2. When training a soft-margin SVM, the noisier is the data the larger should be set the value of hyperparameter C.
3. Hard-margin SVMs can be successfully applied also to datasets that appear to be not linearly separable.
4. The aim of the Kernel Trick is to limit the computational cost of the SVM training on datasets with a very large number of samples.

Exercise 6.15

The client you are working for, Apple & Co., asked you to classify the quality of some fruits (i.e., 1-st quality and 2-nd quality) by basing on their characteristics (i.e., color and weight). You decided to use a linear SVM to solve the problem. After some time, the same client asks you to provide new solutions to improve the capabilities of the classifier you proposed. Comment the following options and tell if they are promising for increasing the testing performance (accuracy) of the SVM.

1. Enhance the training set by getting data points whose values of the input are far from the boundary of the current SVM.
2. Buy a new server in order to be able to apply a kernel on the previous SVM.
3. Enhance the training set by using new data whose input are near to the margins of the current SVM.

4. Introduce new input variables (e.g., diameter, density) and train the SVM on a new dataset containing this information.

Exercise 6.16

Consider the linear two-class SVM classifier defined by the parameters $w = [2\ 1]$, $b = 1$. Answer the following questions providing adequate motivations.

- Is the point $x_1 = [-2\ 4]$ a support vector?
- Give an example of a point which is on the boundary of the SVM.
- How the point $x_2 = [3\ \ -1]$ is classified according to the trained SVM?
- Assume to collect a new sample $x_3 = [-1\ 2]$ in the negative class, do we need to retrain the SVM?

Exercise 6.17

After training a logistic regression classifier with gradient descent on a given dataset, you find that it does not achieve the desired performance on the training set, nor the cross validation one.

Which of the following might be a promising step to take?

1. Use an SVM with a Gaussian Kernel.
2. Introduce a regularization term.
3. Add features by basing on the problem characteristics.
4. Use an SVM with a linear kernel, without introducing new features.

* Exercise 6.18

Derive the dual formulation from the primal SVM minimization problem with soft margins.

Exercise 6.19

Tell if the following statements about SVM are true or false. Motivate your answers.

1. There exists a closed form solution to provide the optimal weights of the SVM.
2. There exists a unique solution for the problem of optimizing the weights of the SVM.

3. Since they have the same activation function, it is equivalent to train an SVM and a perceptron, if we have a linearly separable dataset.

4. The boundary of a nonlinear kernel SVM is linear in a specific high-dimensional feature space.

Exercise 6.20

Tell which of the following methods is a parametric method and which is not. Motivate your answers.

1. Gaussian Processes;

2. Logistic Regression;

3. Ridge Regression;

4. K-Nearest Neighbors.

Exercise 6.21

Tell if the following statements about parametric and non-parametric methods are true or false. Motivate your answers.

1. To address a classification task on a large dataset of low-dimensional points, it is usually better to employ a non-parametric method than a parametric one.

2. When a regression task requires to provide real-time predictions, it is in general a good idea to train a non-parametric method.

3. Non-parametric methods are generally less affected by the curse of dimensionality than parametric methods.

4. The Bayesian linear regression is a non-parametric method.