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.

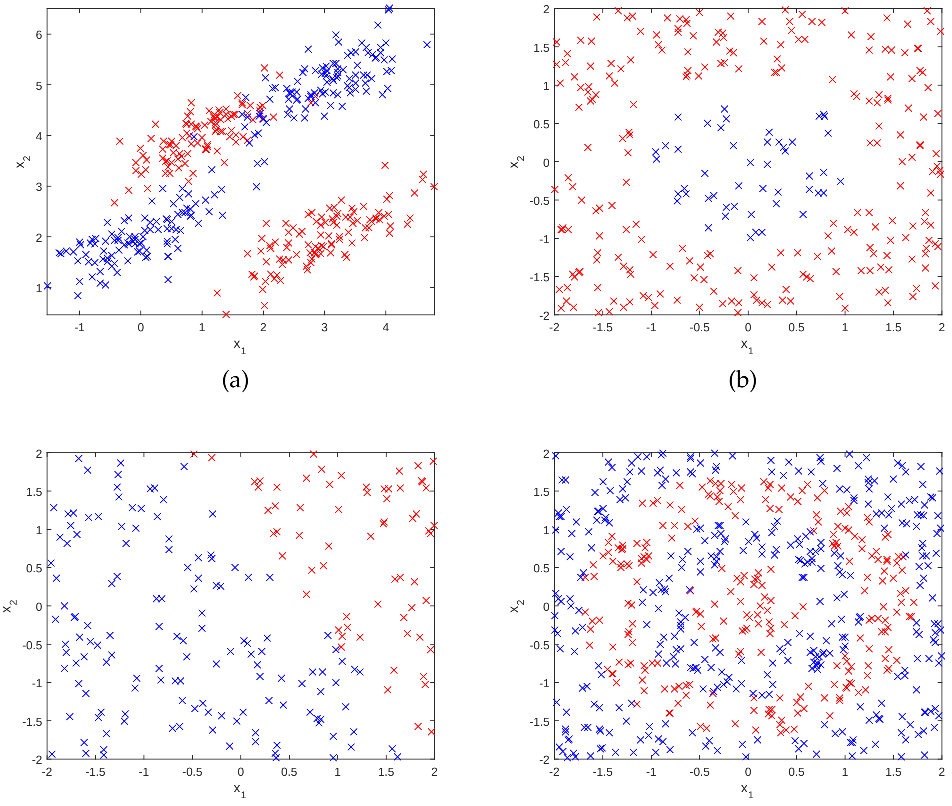
Exercise 6.2

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.

Exercise 6.3

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



(a)

(b)

(c)

(d)

Figure 6.1: Different datasets.

the use of kernels?

* Exercise 6.4

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 6.5

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

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. Which kind of covariance kernel would you use? Provide the analytic form of the kernel and motivate why you would choose it.

There exist other techniques which are able to handle this problem? Are there any drawbacks in doing so?

Why you should not consider such a model in the case you have the information that each dimension is equivalent to the others.

Exercise 6.7

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.

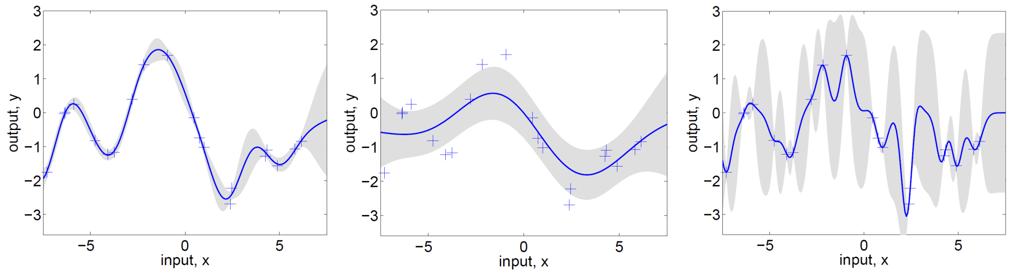
Exercise 6.8

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 6.9

Consider the following formulation of a SVM:

- Hypothesis space: $y_n = f(\mathbf{x}_n, \mathbf{w}) = sign\left(\mathbf{w}^T\mathbf{x}_n + b\right)$;
- Loss measure: $L(X, \mathbf{w}) := ||\mathbf{w}||^2 + C\sum_i \zeta_i \qquad$ s.t. $t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \zeta_i \ \forall n$,

where ζ_i is the violations of the margin provided by sample $\mathbf{x}_i$. Answer the following questions about SVM. Provide adequate explanation for your answer.

1. If we increase the regularization parameter C in an SVM, how do you expect the margin to behave? Do they become thinner or thicker?
2. If no linear boundary can perfectly classify all the training data, this means we need to use a feature expansion. True or false?
3. The computational effort required to solve a kernel support vector machine becomes greater and greater as the dimension of the basis increases. True or false?
4. Suppose that after our computer works for an hour to fit an SVM on a large data set, we notice that x_4, the feature vector for the fourth example, was recorded incorrectly, i.e., we use $\hat{x}_4$ instead of x_4 to train our model. However, your co-worker notices that the pair $(\hat{x}_4 \ y_4)$ did not turn out to be a support point in the

original fit. He says there is no need to train again the SVM on the corrected data set, because changing the value of a non–support point can't possibly change the fit. True or false?

Exercise 6.10

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 6.11

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 6.12

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 6.13

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 6.14

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

Exercise 6.15

What's the black magic of SVMs? More specifically, which parameters we can tune to better fit a specific classification task with a SVM?

Exercise 6.16

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 6.17

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.