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.

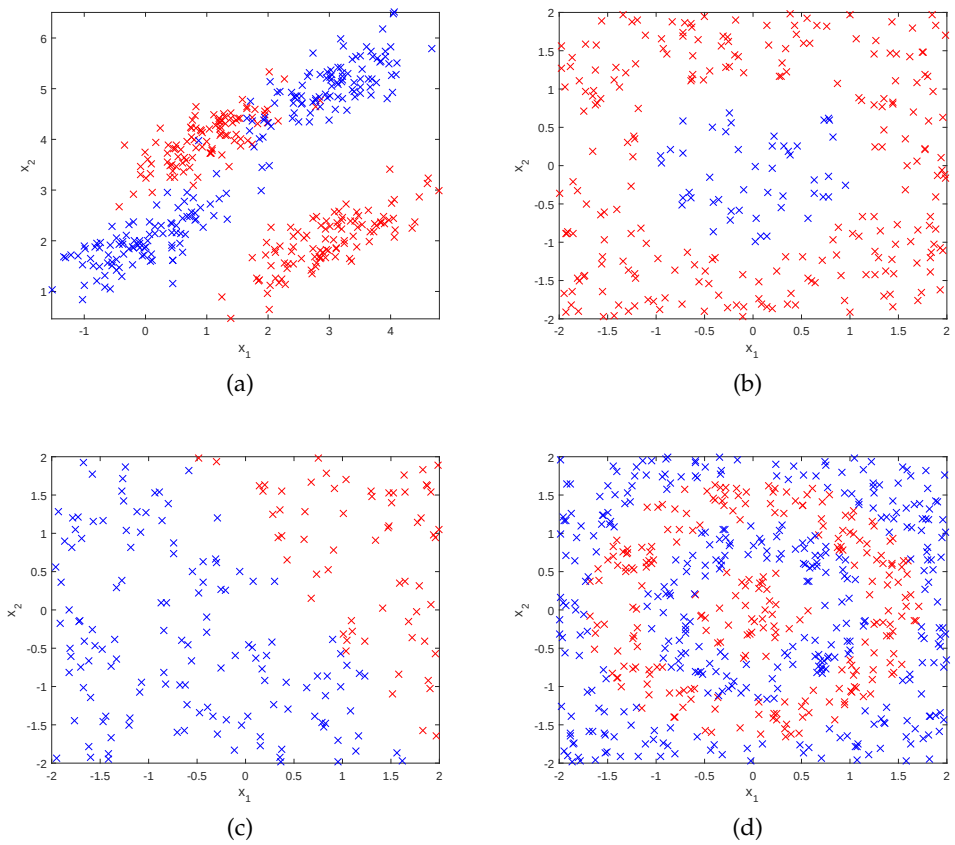


Figure 6.1: Different datasets.

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.

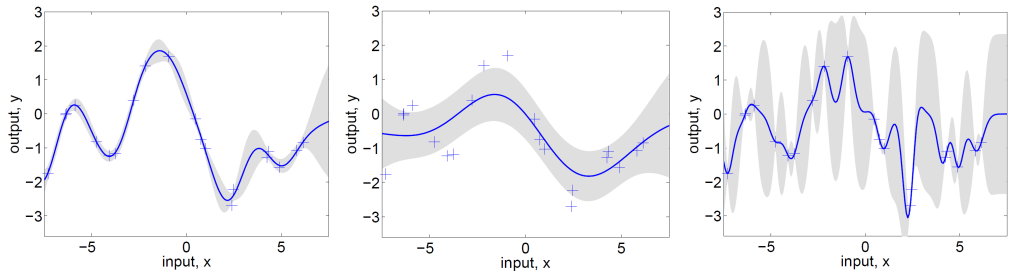
5. As in linear models, we are considering different uncertainty in each point of the input space x.

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 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.

Solutions

Answer of exercise 6.1

1. FALSE: if we add unnecessary features we are increasing the variance of the considered model without having any benefit for decreasing the model bias.

2. TRUE/FALSE: it depends if the approach is parametric or non–parametric. In the first case the training generally increase in computational time. If we consider a non–parametric method, which does not require training, we do not require any computation.

3. TRUE: To compute the prediction of a new samples using more features does require a larger computational time (both for parametric and non–parametric methods), unless you are resorting to the kernel trick, which might keep the computational cost dependent only on the dataset dimension N.

4. TRUE: Usually the choice of features to be added to your model requires a priori information on the problem.

5. FALSE: If you resort to the kernel trick you project your data into a higher dimensional space without requiring to choose the specific features. It is still true that the use of the proper kernel for a specific problem might give better generalization capabilities to your learner.

Answer of exercise 6.2

1. In principle, one should define a metric in the colour space. As long as we are able to define it, we can apply kernels. The other possibility is to transform the colours into binary vectors and apply the kernel to that transformation. Generally, Gaussian kernels are not so effective on binary vectors, therefore one should choose a different shape for the kernel.

2. The same holds for graphs. We need to have a metric to understand how two graphs are similar to each other. For instance, one might use the minimum number of operations needed to transform a graph into the other.

3. The kernel does not increase the number of operations one should perform (computing a distance should be linear in the number of features), while we need to store data about all/most of the training set, therefore we might need a larger hard drive to store it.

4. No: if we have a priori information on the dataset, we should use them and project into a specifically crafted feature space. Yes: if the space in which we

are projecting is an infinite/high dimensional one, we might try to use the kernel trick not to work on large feature vectors.

Answer of exercise 6.4

In Figure 6.1a we have clearly a structure in the data, but it is difficult to evidence a set of features s.t. the classes are linearly separable. In this specific case we might consider the use of Gaussian kernels, since each class seems to be composed by a set of Gaussian in the original input space.

In Figure 6.1b it is possible to find a regularity of the two classes by considering the radius of a circle centred in $(0, 0)$. Thus a transformation of polar coordinates might do the trick. In this case it is better to use the information we extracted from the data then use an arbitrarily complex feature space induced by kernels.

In Figure 6.1c we have a linear separating hyperplane in the feature space. We should use some linear techniques directly applied in this space.

In Figure 6.1d there is a clear structure in the problem. If you are able to find a suitable feature space one might use it. Another simple option is to resort to kernels and "hope" that the problem is linearly separable in this new feature space.

Answer of exercise 6.5

According to *Mercer's theorem*:

Theorem 1. *Any continuous, symmetric, positive semi–definite kernel function $k(x, y)$ can be expressed as a dot product in a high–dimensional space.*

Thus:

1. TRUE it is symmetric and continuous, moreover, given a set of vectors $x_1, \dots x_n$ and a generic vector $z \in \mathbb{R}^d, z = (z_1, \dots z_n)$ we have:

$$\begin{aligned} z^\top K_n z &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j x_i^\top x_j = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \sum_{h=1}^{n} x_{i,h} x_{j,h} \\ &= \sum_{h=1}^{n}\left(\sum_{i=1}^{n} z_i x_{i,h} \sum_{j=1}^{n} z_j x_{j,h}\right) = \sum_{h=1}^{n}\left(\sum_{i=1}^{n} z_i x_{i,h}\right)^2 \geq 0, \end{aligned}$$

   where $(K_n)_{i,j} = x_i^\top x_j$ is the Gram matrix, and $x_{i,h}$ is the h-th element of the vector x_i. ;

2. TRUE since it is a composition (sum and square) of a valid kernel;

3. TRUE if $c > 0$ for the rules on the composition of valid kernels;
4. FALSE since the kernel function is not symmetric;
5. TRUE since A is semidefinite positive we can define B = $A^{1/2}$ and we have that:

$$z^\top K_n z = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j (x_i B)^\top (B x_j),$$

and we can repeat the process we used for the linear kernel;

6. FALSE if $x, y \in \mathbb{R}$, TRUE if $x, y \in [2k\pi, (2k+1)\pi], k \in \mathbb{N}$.

Answer of exercise 6.6

1. TRUE. By definition, selecting $\phi(\mathbf{x}) = \mathbf{x} + \mathbf{1}$, we have $k_1(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$.
2. FALSE. $k_2(\mathbf{x}, \mathbf{y})$ is not symmetric.
3. TRUE. Since we are considering the same transformation $\cos(\cdot)$ applied to both arguments, then $(\cdot)^3$ is a polynomial transformation with non-negative coefficients, and k_1 is a kernel.
4. TRUE. With simple computations, we obtain:

$$k_4(\mathbf{x}, \mathbf{y}) = \exp\left(-k_2(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{y}, \mathbf{x})\right) = \exp\left(2\mathbf{x}^T\mathbf{y} - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2\right) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|^2\right)$$

that is the Gaussian kernel with $\sigma^2 = \frac{1}{2}$.

Answer of exercise 6.7

1. In the case we have different scales for the different dimensions we might consider a different bandwidth for each dimension. For instance:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \theta_0 \exp\left\{-\sum_h \frac{(x_{ih} - x_{jh})^2}{2\theta_h^2}\right\},$$

where we have $\theta \in \mathbb{R}^{M+1}$ parameter vector, so that the correlation of different dimension influence differently the prediction process.

2. Another possible choice is to perform z-scoring on the input variables, s.t., all the variables have the same range, thus are more likely to have the same behaviour over all the dimensions. With this method we are not sure that the problem of different lengthscale is completely solved.

3. In the case we have prior knowledge that all the dimensions have the same behaviour we could consider a single parameter for the bandwidth, otherwise we are considering an unnecessary complex model for the data we would like to analyse.

Answer of exercise 6.8

1. TRUE/FALSE If we assume a zero variance noise, the posterior shrinks as new data are coming. Otherwise, even with infinite number of points we have that the GP in a point x has some variance.

2. FALSE As usual in ML we are assured to converge if the prior is either uninformative or generic enough, i.e., if properly integrates some prior knowledge without biasing the learning process too much. There are cases in which a wrong prior prevents from converging to the true function (e.g., null prior probability on the true parameters).

3. FALSE By considering, for instance, a logit function in conjunction with the GP we can also tackle binary classification problems (the output in this case is a probability).

4. TRUE Since the kernel usually influence more the nearest points and is does not induce any change in the behaviour of the far points (which are mainly determined by the process noise).

5. FALSE In linear models we are using a single variance for the noise throughout the entire input space x. This is also true for the GPs. The different uncertainty is determined by the covariance structure.

Answer of exercise 6.9

By looking at the bandwidth we can associate each figure to each parameter vectors. Indeed, the more we have larger bandwidth the more the process is smooth (i.e., the frequency over the input space of th GP is lower). Thus, the correct correspondence is:

- $1 \rightarrow (a)$
- $2 \rightarrow (c)$
- $3 \rightarrow (b)$

Answer of exercise 6.10

1. TRUE: they provide a way of computing the expected value and the variance of each point in the input space, which could be different from point to point.

2. FALSE: in their original formulation the GP have been designed for regression, though they can be used also for other tasks.

3. TRUE: from the definition of GP.

4. TRUE: since a GP bases its shape from the existing data and from the covariance structure, if we are in a region of the space where we do not have samples, they will provide an uncertain prediction.

Answer of exercise 6.11

1. FALSE: they require to store the gram matrix whose dimension depends on the number of samples.

2. TRUE: it requires only the gram matrix and the computation of the kernel on the new point.

3. TRUE: the uncertainty we have around the sampled points decreases as we get more and more samples.

4. TRUE: indeed, it requires the inversion of the gram matrix which has N^3 computational cost.

Answer of exercise 6.13

1. TRUE The boundary is linear in the feature space as specified since it has the same shape as the one we analyse when talking about the perceptron.

2. FALSE The use of the kernel is completely independent on the method used for classifying multiple classes. Sometimes the fact that the boundaries between many classes is not jointly linear suggests to use kernels.

3. TRUE Since $\max_x e^{-x^2} = 1$.

4. FALSE The parameter C regulates somehow the thickness of the box between the margins and thus influence the boundary parameters $\mathbf{w}$ too.

Answer of exercise 6.14

1. FALSE: SVM are sparse model and the prediction cost scales with the number of support vectors.

2. FALSE: the larger is C the lower is the bias (and conversely the higher the variance). This is exactly the opposite of what we should expect to do on a noisy dataset.

3. TRUE: applying the kernel function we could find the problem to bel linearly separable in the feature space.

4. FALSE: the computational cost of training SVM is unfortunately cubic in the number of training samples. Kernel trick allows to limit the cost of computing a large number of features.

Answer of exercise 6.15

1. FALSE: points which are far from the separating hyperplane are not likely to change the result of the SVM training process.

2. FALSE: the use of kernel does not require more computational power than the linear one.

3. TRUE: the points are likely to become the new support vectors and modify the SVM separating surface.

4. TRUE: as long as they are meaningful for the problem it is always a good idea to use new input variables.

Answer of exercise 6.16

- A point is a support vector if $|w^T x + b| \leq 1$, thus:

$$\left|w^T x + b\right| = |-4 + 4 + 1| = 1$$

  meaning that x_1 is a support vector.

- A point on the boundary has to satisfy $w^T x + b = 0$ thus by considering $x_{11} = 0$:

$$2 \cdot 0 + 1 \cdot x_{22} + 1 = 0 \rightarrow x_{22} = -1$$

  thus $x = [0 \ -1]$ is on the boundary.

- A point is classified either in the positive class or in the negative one if $w^T x + b$ is positive or negative, respectively, thus:

$$w^T x_2 + b = 2 \cdot 3 - 1 \cdot 1 + 1 = 6$$

  which means that the point it classified in the positive class.

- The point is misclassified by the current model, as it is

$$w^T x_3 + b = -2 + 2 + 1 = 1$$

  which means that x_3 would be a support vector, thus we need to retrain the model.

Answer of exercise 6.17

1. OK In this case we are considering a new feature space where the classification task might be viable and the amount of samples we have allows us to expand our features space further.

2. NO In this case the regularization would even increase the error on the training set.

3. OK In higher dimensional space there might be the chance that the classes are linearly separable and thus that logistic regression is performing well. This clearly requires you to have a priori information about the considered problem.

4. NO With different optimization techniques and target, both logistic regression and SVM are finding a separating hyperplane. If has not been found by logistic regression, we have no chance that the SVM is able to find this separating surface.

Answer of exercise 6.18

See Page 410 of **Friedman, Jerome, Trevor Hastie, and Robert Tibshirani**. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

Answer of exercise 6.19

1. FALSE The problem of finding the optimal parameter requires to transform the primal optimization problem into the dual one and apply an iterative procedure (i.e., SMO) to find the optimal weights.

2. TRUE The objective function of the SVM is convex, therefore it has a unique minimum.

3. FALSE Even if they have the $sign(\cdot)$ activation function, they are optimizing different objectives. The perceptron is minimizing the misclassification error, while the SVM is maximizing the intraclass margin. This holds also for linearly separable datasets.

4. TRUE The idea of using a kernel is based on the idea that expanding the features into a complex space it is more likely the data can be all correctly classified by a straight boundary (i.e., a hyperplane).

Answer of exercise 6.20

1. NONPARAMETRIC: it require to store the initial data and compute the value of the kernel to provide a prediction;

2. PARAMETRIC: the final result of the optimization problem is the parameter vector of the weights $\mathbf{w}$. The training set can be discarded once we have this vector.

3. PARAMETRIC and NONPARAMETRIC: during the course we saw that this approach can be casted in both ways depending if the number of parameterts is large or small.

4. NONPARAMETRIC: the provide a prediction one needs the entire dataset. No data from the training can be discarded.

Answer of exercise 6.21

1. FALSE, non-parametric methods are usually convenient when data are high-dimensional, whereas parametric methods can better handle very large low-dimensional datasets.

2. FALSE, non-parametric methods (e.g., k-nearest neighbors) are usually slower in prediction than parametric methods (e.g., linear regression).

3. TRUE, non-parametric methods can generally deal with an exponentially-growing feature space through the kernel trick.

4. FALSE, Bayesian linear regression is a parametric method, in which the parameters are learned from data through a Bayesian approach.