Exercises for the course
Machine Learning 2
Summer semester 2021

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 8

Exercise 1: One-Class SVM (5 + 5 + 20 + 10 + 10 P)

The one-class SVM is given by the minimization problem:

$$\min_{\boldsymbol{w},\rho,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{N\nu}\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} \quad \forall_{i=1}^{N} : \langle \phi(\boldsymbol{x}_i), \boldsymbol{w}\rangle \geq \rho - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the training data and $\phi(\boldsymbol{x}_i) \in \mathbb{R}^d$ is a feature space representation.

(a) *Show* that strong duality holds (i.e. verify the Slater's conditions).

(b) *Write* the Lagrange function associated to this optimization problem.

(c) *Show* the dual program for the one-class SVM is given by:

$$\max_{\boldsymbol{\alpha}} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i = 1 \quad \text{and} \quad \forall_{i=1}^{N} : \ 0 \leq \alpha_i \leq \frac{1}{N\nu}$$

(d) *Show* that the problem can be equivalently rewritten in canonical matrix form as:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

$$\text{s.t.} \quad \mathbf{1}^\top\boldsymbol{\alpha} = 1 \quad \text{and} \quad \begin{pmatrix} -I \\ I \end{pmatrix}\boldsymbol{\alpha} \preceq \begin{pmatrix} \mathbf{0} \\ \mathbf{1}/N\nu \end{pmatrix}$$

where K is the Gram matrix whose elements are defined as $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

(e) The decision rule in the primal for classifying a point as an outlier is given by:

$$\langle \phi(\boldsymbol{x}), \boldsymbol{w}\rangle < \rho$$

Also, one can verify that for any data point $\boldsymbol{x}_i$ whose associated dual variable satisfies the strict inequalities $0 < \alpha_i < \frac{1}{N\nu}$, and calling one such point a support vector $\boldsymbol{x}_{\text{SV}}$, the following equality holds:

$$\langle \phi(\boldsymbol{x}_{\text{SV}}), \boldsymbol{w}\rangle = \rho$$

Show that the outlier detection rule can be expressed as:

$$\sum_{i=1}^{N}\alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) < \sum_{i=1}^{N}\alpha_i k(\boldsymbol{x}_{\text{SV}}, \boldsymbol{x}_i)$$

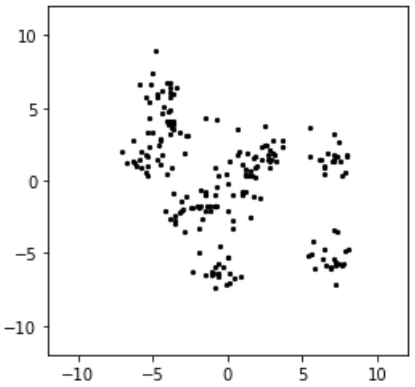
Exercise 2: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Implementing Anomaly Detection Models

In this exercise sheet, several kernel-based anomaly detection models will be implemented and their behavior compared on a simple two-dimensional dataset. The following code builds a dataset generated as a mixture of several Gaussian blobs.

In [1]:
```
import sklearn.datasets
import sklearn.metrics
import numpy
import matplotlib
from matplotlib import pyplot as plt
%matplotlib inline
import utils

X = sklearn.datasets.make_blobs(n_samples=200,centers=10,random_state=2)[0]
X = X - X.mean(axis=0)
X = X / X.std() * 4.0

utils.plot(X,None)
```



Kernel Density Estimation (10 P)

The first anomaly detection model is based on kernel density estimation (KDE). KDE builds the function

$$f(x) = \frac{1}{N}\sum_{n=1}^{N} k(x, x_n)$$

where the output forms here an unnormalized probability density function. Note that if only interested in producing an ordering of points from least to most outlier, we don't need to normalize $f(x)$. However, because $f(x)$ is more a measure of inlierness than outlierness, we can define the outlier score $o(x)$ as a decreasing function of $f(x)$ and also make sure the function goes to infinity for very remote data points. This can be achieved with the scoring function:

$$o(x) = -\log(f(x))$$

We now would like to implement KDE using an interface similar to how ML algorithms are provided in scikit-learn, in particular, by defining a class that implements a `fit` function for training based on some training data `X` and a `predict` function for computing the prediction for a new set of points `X`. The KDE class is initialized with a kernel function (typically a Gaussian kernel). Its functions for training and predicting are incomplete.

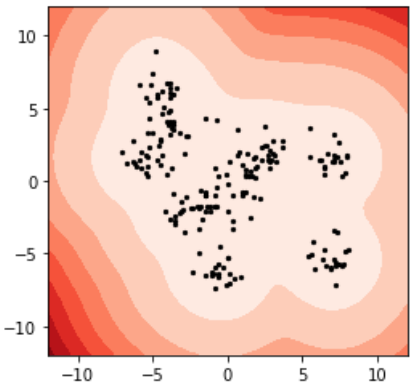
Task:

- **Implement the functions `fit(self,X)` and `predict(self,X)` of the class `KDE`.**

In [2]:
```
class KDE:

    def __init__(self,kernel):
        self.kernel = kernel

    def fit(self,X):

        # ----------------------------------
        # TODO: replace by your code
        # ----------------------------------
        import solution
        solution.kde_fit(self,X)
        # ----------------------------------

        return self

    def predict(self,X):

        # ----------------------------------
        # TODO: replace by your code
        # ----------------------------------
        import solution
        o = solution.kde_predict(self,X)
        # ----------------------------------

        return o
```

The KDE model can now be tested on our two-dimensional data. The code below passes to the KDE model a Gaussian kernel of scale $\gamma = 0.25$ (i.e. the bandwidth is slightly larger than for the default Gaussian kernel), train the model on the Gaussian blobs data, and apply the model to a grid dataset for the purpose of building a contour plot.

In [3]:
```
kernel = lambda x,y: sklearn.metrics.pairwise.rbf_kernel(x,y,gamma=0.25)
utils.plot(X,KDE(kernel).fit(X).predict(utils.Xgrid))
```



We observe that model behaves as expected, i.e. the regions outside the data are highlighted in red, which corresponds to high outlier scores.

Uncentered Kernel PCA Anomaly Detection (15 P)

Another model for anomaly detection is based on Kernel PCA. Here, we consider an uncentered version of Kernel PCA where we do not subtract the mean of the data in feature space. Because it is not possible to compute exactly the eigenvectors from finite data, we resort to an empirical approximation based on the Gram matrix:

$$[K]_{nn'} = k(x_n, x_{n'})$$

and diagonalizing it to get empirical eigenvectors and eigenvalues:

$$K = U \Lambda U^\top$$

The matrix Λ is diagonal and contains all eigenvalues $\lambda, \ldots, \lambda_N$ sorted in descending order. The columns of the matrix U are the corresponding eigenvectors. For the training data, projection of the nth data point on the ith principal component is readily given by

$$\mathrm{proj}_i(x_n) = U_{n,i} \cdot \lambda_i^{0.5}$$

For new data points $x \in \mathbb{R}^d$, such projection is not readily available and we can resort instead to the following interpolation scheme:

$$\mathrm{proj}_i(x) = k(x, X) \cdot U_{:,i} \cdot \lambda_i^{-0.5}$$

The latter produces equivalent results for points $(x_n)_n$ in the dataset but it generalizes the projection to any other point $x \in \mathbb{R}^d$. Once the data has been projected on the principal components, the outlier score can be computed as:

$$o(x) = k(x, x) - \sum_{i=1}^{a} \big(\mathrm{proj}_i(x)\big)^2$$

An incomplete version of uncentered kernel PCA anomaly detection is given below. Like for KDE, it receives a kernel as input, but one also needs to needs to specify the number of dimensions used in the Kernel PCA model.

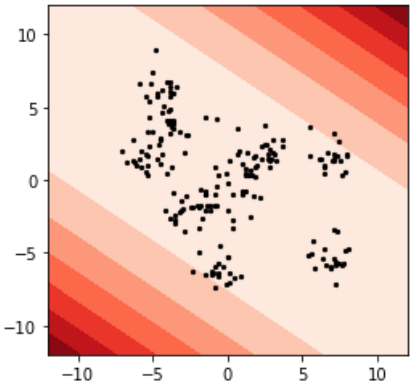
Task:

- **Implement the functions `fit(self,X)` and `predict(self,X)` of the class `UKPCA`.**

In [4]:
```
class UKPCA:

    def __init__(self,kernel,dims):
        self.kernel = kernel
        self.dims = dims

    def fit(self,X):

        # ----------------------------------
        # TODO: replace by your code
        # ----------------------------------
        import solution
        solution.ukpca_fit(self,X)
        # ----------------------------------

        return self

    def predict(self,X):

        # ----------------------------------
        # TODO: replace by your code
        # ----------------------------------
        import solution
        o = solution.ukpca_predict(self,X)
        # ----------------------------------

        return o
```

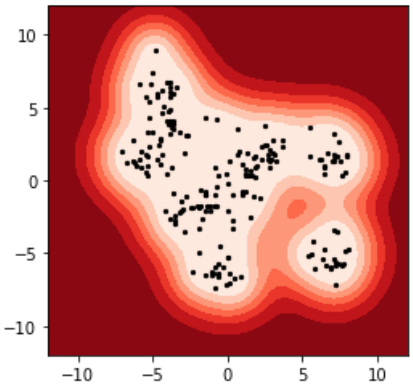
The kernel PCA approach can now be tested. We first consider a kPCA model with a linear kernel and where we retain only the first principal component.

In [5]:
```
kernel = sklearn.metrics.pairwise.linear_kernel
utils.plot(X,UKPCA(kernel,1).fit(X).predict(utils.Xgrid))
```



The outlier score grows along the second principal component (the one with least variance). We now consider instead a Gaussian kernel (of slightly larger bandwidth than the one used for KDE) and build a the outlier function from a KPCA model containing 25 principal components.

In [6]:
```
kernel = lambda x,y: sklearn.metrics.pairwise.rbf_kernel(x,y,gamma=0.1)
utils.plot(X,UKPCA(kernel,25).fit(X).predict(utils.Xgrid))
```



Here, we observe that the outlier model much more closely follows the shape of the data distribution. However, we also observe that it saturates away from the data, which does not reflect the true degree of outlierness.

One-Class SVM (25 P)

The one-class SVM is another approach to anomaly detection that aims to build some envelope that contains the inliner data and that separates it from outlier data. In its dual form, it consists of solving the constrained optimization problem:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

subject to

$$\mathbf{1}^\top \boldsymbol{\alpha} = 1 \quad \text{and} \quad \begin{pmatrix} -I \\ I \end{pmatrix} \boldsymbol{\alpha} \preceq \begin{pmatrix} \mathbf{0} \\ \mathbf{1}/N\nu \end{pmatrix}$$

To solve this optimization problem, we can use the quadratic solver provided as part of `cvxopt` and the interface of which is shown below:

cvxopt.solvers.qp(*P, q* [, *G, h* [, *A, b* [, *solver* [, *initvals*]]]])

Solves the pair of primal and dual convex quadratic programs

$$\begin{array}{ll} \text{minimize} & (1/2)x^T P x + q^T x \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

Once the solution has been found, the output score can be computed as $f(x) = \sum_i \alpha_i k(x, x_i)$. Similarly to the outlier scores we have computed for KDE, we can build a transformation

$$o(x) = -\log \frac{\sum_i \alpha_i k(x, x_i)}{\sum_i \alpha_i k(x_{\text{SSV}}, x_i)}$$

where x_{SSV} is any 'strict' support vector (they can be identified as implementing the box constraints above with strict inequalities). With this transformation the equation $o(x) = 0$ also gives the OC-SVM decision boundary.

Task:

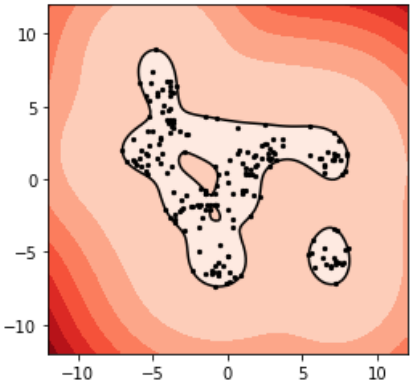
- **Implement the functions `fit(self,X)` and `predict(self,X)` of the class `OCSVM`.**

In [7]:
```
import cvxopt
import cvxopt.solvers

class OCSVM:

    def __init__(self,kernel,nu):
        self.kernel = kernel
        self.nu = nu

    def fit(self,X):

        # ---------------------------------
        # TODO: replace by your code
        # ---------------------------------
        import solution
        solution.ocsvm_fit(self,X)
        # ---------------------------------

        return self

    def predict(self,X):

        # ---------------------------------
        # TODO: replace by your code
        # ---------------------------------
        import solution
        o = solution.ocsvm_predict(self,X)
        # ---------------------------------

        return o
```

The OC-SVM can now be tested on the 2d dataset. Here, we first consider the case where $\nu = 0.0001$, which corresponds to implementing a hard envelope (with no points outside of it).

In [8]:
```
kernel = lambda x,y: sklearn.metrics.pairwise.rbf_kernel(x,y,gamma=0.1)
utils.plot(X,OCSVM(kernel,0.0001).fit(X).predict(utils.Xgrid),boundary=True)
```

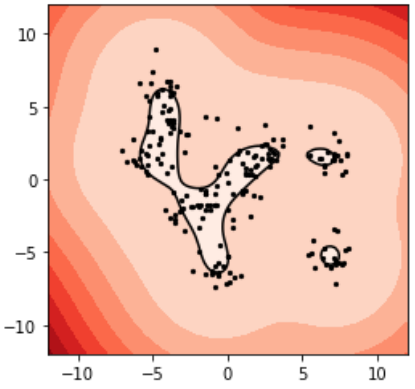
```
     pcost       dcost       gap    pres   dres
 0:  5.9756e-02 -1.0097e+04  1e+04  1e-13  2e-13
 1:  5.9751e-02 -1.0853e+02  1e+02  3e-15  4e-13
 2:  5.9521e-02 -4.9829e+00  5e+00  2e-16  2e-14
 3:  7.0927e-02 -4.2701e+00  4e+00  4e-16  1e-14
 4:  7.4758e-02 -3.5860e+00  4e+00  4e-16  2e-14
 5:  6.8449e-02 -1.8468e-01  3e-01  2e-16  2e-15
 6:  6.2819e-02 -1.0865e-01  2e-01  2e-16  9e-16
 7:  5.9641e-02  1.9806e-02  4e-02  3e-16  8e-16
 8:  5.5603e-02  3.9579e-02  2e-02  7e-16  7e-16
 9:  5.4303e-02  4.7196e-02  7e-03  6e-16  7e-16
10:  5.3536e-02  5.1293e-02  2e-03  2e-16  6e-16
11:  5.3182e-02  5.2474e-02  7e-04  4e-16  7e-16
12:  5.3067e-02  5.2747e-02  3e-04  4e-16  7e-16
13:  5.2983e-02  5.2925e-02  6e-05  4e-16  6e-16
14:  5.2968e-02  5.2951e-02  2e-05  4e-16  6e-16
15:  5.2963e-02  5.2960e-02  3e-06  2e-16  7e-16
16:  5.2962e-02  5.2961e-02  3e-07  2e-16  7e-16
17:  5.2961e-02  5.2961e-02  7e-09  2e-16  7e-16
Optimal solution found.
```



We observe that all points are indeed either contained in the envelope or at the border of it. We can now test the OC-SVM with a larger parameter ν, here, $\nu = 0.1$ and run the code again:

In [9]:
```
kernel = lambda x,y: sklearn.metrics.pairwise.rbf_kernel(x,y,gamma=0.1)
utils.plot(X,OCSVM(kernel,0.5).fit(X).predict(utils.Xgrid),boundary=True)
```

```
     pcost       dcost       gap    pres   dres
 0:  5.9756e-02 -1.9792e+00  4e+02  2e+01  2e-15
 1:  6.6548e-02 -1.9549e+00  6e+00  2e-01  4e-15
 2:  7.2113e-02 -8.7545e-01  9e-01  4e-16  3e-15
 3:  7.0349e-02  1.9189e-02  5e-02  3e-17  1e-15
 4:  6.5181e-02  5.4344e-02  1e-02  6e-16  9e-16
 5:  6.3039e-02  5.9564e-02  3e-03  9e-17  7e-16
 6:  6.2249e-02  6.0824e-02  1e-03  2e-17  6e-16
 7:  6.1873e-02  6.1403e-02  5e-04  1e-16  6e-16
 8:  6.1728e-02  6.1573e-02  2e-04  3e-16  7e-16
 9:  6.1671e-02  6.1646e-02  3e-05  1e-16  8e-16
10:  6.1661e-02  6.1659e-02  2e-06  5e-16  7e-16
11:  6.1660e-02  6.1660e-02  4e-08  4e-16  7e-16
Optimal solution found.
```



This time, not all data points are contained in the envelope, and some of them are therefore classified by the model as outlier.

Exercise Sheet 8

1 One-Class SVM

a) $-\xi_i \geq 0$ if we set $\xi_i \geq 0$

$-\langle \phi(x_i), w \rangle \geq \rho - \xi_i$ can be achieved by setting ξ_i very large ✓

b) $\mathcal{L}(w, \rho, \xi, \lambda_1, \lambda_2) = \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu N}\sum_{i=1}^{N} \xi_i$

$+ \sum_i \lambda_{1,i} \underbrace{(\rho - \xi_i - \langle \phi(x_i), w \rangle)}_{\leq 0} + \sum_i \lambda_{2,i} \underbrace{(-\xi_i)}_{\leq 0}$

c) $\max_{\alpha \geq 0, \beta \geq 0} \min_{w, \rho, \xi} \mathcal{L}(\cdot)$ $\qquad \lambda_1 \to \alpha, \lambda_2 \to \beta$

$\frac{\partial \mathcal{L}(\cdot)}{\partial w} = w - \sum_i \alpha_i \phi(x_i) \overset{!}{=} 0 \iff w = \sum_i \alpha_i \phi(x_i)$

$\frac{\partial \mathcal{L}(\cdot)}{\partial \rho} = -1 + \sum_i \alpha_i \overset{!}{=} 0 \iff \sum_i \alpha_i \overset{!}{=} 1$

$\frac{\partial \mathcal{L}(\cdot)}{\partial \xi_i} = \frac{1}{\nu N} - \alpha_i - \beta_i = 0$

$\max_{\alpha \geq 0, \beta \geq 0} \frac{1}{2}\left\| \sum_i \alpha_i \phi(x_i) \right\|^2 - \sum_i \alpha_i \left\langle \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle$

$= \max_{\alpha \geq 0} -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \underbrace{\phi(x_i)^T \phi(x_j)}_{k(x_i, x_j)}$

$\sum_i \alpha_i = 1, \quad \alpha_i \geq 0, \quad \frac{1}{\nu N} - \alpha_i - \beta_i = 0 \iff \frac{1}{\nu N} \geq \alpha_i$

$\to \quad 0 \leq \alpha_i \leq \frac{1}{\nu N}$

d) $\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(x_i,x_j) \rightarrow \min_{\alpha} \frac{1}{2}\sum_{ij}\alpha_i\alpha_j \underbrace{k(x_i,x_j)}_{K_{ij}}$

$\underbrace{\qquad\qquad}_{\alpha^T K \alpha}$

$\sum_{i}^{N}\alpha_i = 1 \Leftrightarrow 1^T\alpha = 1$

$0 \leq \alpha_i \leq \frac{1}{NC} \rightarrow \begin{pmatrix} -I \\ I \end{pmatrix}\alpha \leq \begin{pmatrix} 0 \\ \frac{1}{NC} \end{pmatrix}$

e) $\langle \phi(x_i), w\rangle < \langle \phi(x_{sv}), w\rangle$

$w = \sum_i \alpha_i \phi(x_i)$

$\langle \phi(x_i), \sum_i \alpha_i \phi(x_i)\rangle < \langle \phi(x_{sv}), \sum_i \alpha_i \phi(x_i)\rangle$

$= \sum_i \alpha \underbrace{\langle \phi(x_i), \phi(x_i)\rangle}_{k(x_i,x_i)} < \sum_i \alpha_i \underbrace{\langle \phi(x_{sv}), \phi(x_i)\rangle}_{k(x_{sv},x_i)}$

$= \sum_i \alpha_i k(x_i,x_i) < \sum_i \alpha_i k(x_{sv},x_i)$