SVM and Optimization

4ème année

G. Gasso

1 Hand made

Let consider the following samples $(\mathbf{x}_i \in \mathbb{R}^2, y_i)$:

Point	Abscissa	ordinate	label y_i
$\mathbf{x}_1$	1	1	-1
$\mathbf{x}_2$	2	0	-1
$\mathbf{x}_3$	-1	-1	1
$\mathbf{x}_4$	1	-2	1

To classify the samples we consider a SVM problem with decision function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$. The problem to be solved is formulated as

$$\begin{array}{ll} \min_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \forall i = 1, \cdots, 4 \end{array}$$

1. Write the dual problem related to this SVM. Formulate it in the matrix form

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha} \in \mathbb{R}^n} & \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{G} \boldsymbol{\alpha} - \boldsymbol{e}^\top \boldsymbol{\alpha} \\ \text{s.t.} & \boldsymbol{z} \leq \boldsymbol{\alpha}, \ \boldsymbol{\alpha}^\top \boldsymbol{y} = 0 \end{array}$$

   where matrix $\mathbf{G}$ and vectors $\boldsymbol{e}$, $\boldsymbol{z}$ and $\boldsymbol{y}$ are to be defined.

2. Let assume that an approximate solution of the dual is given by $\alpha_1 = 0$, $\alpha_2 = 0.4$, $\alpha_3 = 0$ and $\alpha_4 = 0.4$. The intercept b is computed as $b = -0.2$.
   (a) Deduce the value of w from the dual solution.
   (b) Draw the training samples. According to the dual solution emphasize on the figure the support vectors. Justify your response.

3. We are given the validation samples presented below.

Point	Abcisse	Ordonnée	label y_i
$\mathbf{x}_5$	0	1	-1
$\mathbf{x}_6$	0	-1	1
$\mathbf{x}_7$	-1	1	1
$\mathbf{x}_8$	1	-3/4	-1

   Compute the error rate (the relative number of mis-classified samples) on validation.

2 SVM on your own

The goal is to implement a basic SVM solution $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ for linearly separable classification problem using CVXPy. Notice that we do not seek a scalable implementation but a proof of concept. For a scalable SVM solver you should rather consider LibSVM or its wrapper in Scikit Learn for large scale dataset.

Let be a binary classification problem with class $y \in \{-1, 1\}$. Each class is characterized by a normal conditional distribution with mean μ_k and covariance matrix $\mathbf{S}_k$.

1. Generate $n_1 = n_2 = 50$ training samples per class using $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mathbf{S}_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$ and $\mathbf{S}_2 = \mathbf{S}_1$.

```python
import numpy as np
from utility_svm import gen_data_twogaussians_2d

# class 1
n1 = 50
mu1 = np.array([0, 0]); S1 = np.array([[1, 0.5], [0.5, 1]])
# class 2
n2 = n1
mu2 = np.array([5, 5]); S2 = S1

X, Y = gen_data_twogaussians_2d(mu1, S1, mu2, S2, n1, n2)
```

Visualize the samples. Check that the problem is linearly separable (if not generate new data)

```python
import matplotlib.pyplot as plt

# trace des vraies classes
plt.figure(figsize=(8,6))
plt.scatter(X[:,0], X[:, 1], c=Y, cmap="RdYlBu")
plt.colorbar(ticks=np.unique(Y))
```

2. Let solve the dual problem using CVXpy. We will use the dual as expressed in question of the previous exercise.

   (a) Form the matrix $\mathbf{G}$, the vectors e of size n with entries equal to 1, and z a similar vector with all zero entries.

   ```python
   n = X.shape[0]
   e = ...
   z = ...
   ```

   (b) Using CVXpy compute the dual solution $\boldsymbol{\alpha}$ (refer to Constrained Optimization session for a reminder).

   ```python
   import cvxpy as cvx

   lam = 1e-8
   G = G + lam*np.eye(n) # for a better numerical stability
   alpha = cvx.Variable(n)
   obj_dual = cvx.Minimize(....)
   constr_dual = [..., ...]
   dual = cvx.Problem(obj_dual, constr_dual)
   dual.solve(verbose=True)
   alpha = np.squeeze(np.asarray(alpha.value))
   ```

Check the values of $\boldsymbol{\alpha}$ and infer well classified samples, and samples located on the margin.

```
support_vectors = np.where(alpha >= 1e-5)[0]
print("Number of support vectors = {}".format(support_vectors.size))
```

3. Knowing $\boldsymbol{\alpha}$, compute the parameter vector $\mathbf{w}$ of your SVM.
   From the estimated $\mathbf{w}$, compute the bias term b of the SVM. Hint : use the KKT conditions associated to the identified support vectors.
4. Implement your decision rule by writing a function
   `[ypred, fx]` $\leftarrow$ `mysvmpredict(`$\mathbf{w}$`,` b`,` $\mathbf{x}$`)`
   which takes any sample $\mathbf{x}$ as input and outputs the predicted class `ypred` $\in \{-1, 1\}$ and the score `fx` $= \mathbf{w}^\top \mathbf{x} + b$
5. Plot the decision function together with the margin

```
plt.figure(figsize=(8,6))
plt.scatter(X[:,0], X[:, 1], c=Y, cmap="RdYlBu")
plt.colorbar(ticks=np.unique(Y))

# plot the decision frontier and the margin
xx = ...
yy = ...
tmp, f_svm = mysvmpredict ...

cp = plt.contour(xx, yy, f_svm.reshape(n,n), [-1, 0, 1])
plt.clabel(cp, inline=True, fmt="%1.1f", fontsize=10)
```

   Highlight the support vectors on the plot.
6. Can you extend this to a non separable classification problem ?

3 SVM with asymmetric costs

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1,\ldots,n}$ a set of labeled data with $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{X} = \mathbb{R}^d$. We seek a linear SVM model with regularization parameter C_i specific to each sample $\mathbf{x}_i$. The optimization is therefore

$$
\begin{array}{lll}
\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \xi_i} & \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^n C_i \xi_i & \\
\text{s.t.} & y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i & \forall \quad i = 1, \ldots, n \\
 & \xi_i \geq 0 & \forall \quad i = 1, \cdots, n
\end{array}
$$

Here the ξ_i are the slack variables and the C_i denote the regularization parameters (they are fixed by the end-user).

1. Give the lagrangian $\mathcal{L}$ of the problem.
2. Express the KKT stationary condition according to the primal variables $\mathbf{w}, b, \xi_k$ for all $k = 1, n$.
3. Deduce the dual problem. Propose a way to solve the dual problem ?

4. Knowing the dual solution propose a way to compute the intercept b.
5. Let denote as $\mathcal{D}^+ = \{(\mathbf{x}_i, y_i), y_i = 1\}$ and $\mathcal{D}^- = \{(\mathbf{x}_i, y_i), y_i = -1)\}$ respectively the subset of positives and negatives. Now let fixe the following asymmetric costs $C_i = C_+, \ \forall i \in \mathcal{D}^+$ and $C_i = C_-, \ \forall i \in \mathcal{D}^-$.

   Using your previous derivation, propose a way to solve the asymmetric costs SVM. In which application context this problem might be helpful ?

4 Minimum volume data description

Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ be a set of n points of $\mathbb{R}^2$. We want to estimate the minimum volume sphere surrounding the data i.e. the radius and the center of the sphere.

1. Show that this problem can be formulated as

$$\begin{array}{l} \min_{R\in\mathbb{R},\mathbf{c}\in\mathbb{R}^n} R^2 \\ \text{sc. } \|\mathbf{x}_j - \mathbf{c}\|^2 \leq R^2 \quad \forall j \end{array}$$

   with $\mathbf{c}$ the center and R the radius.
2. By assessing the optimality conditions of this problem, show that the center $\mathbf{c}$ can be written as a linear convex combination of the samples $\mathbf{x}_i$.
3. Derive then dual problem formulation.

5 Novelty detection

Let $\{\mathbf{x}_i \in \mathbb{R}^2\}_{i=1,\ldots,n}$ be a set of unsupervised samples. Our objective is to estimate the minimum volume set enclosing as much data as possible while excluding the potential outliers. This amounts to find the smallest sphere with center $\mathbf{c}$ and radius R which is solution of the following problem

$$\begin{array}{lll} & \min_{R,\mathbf{c}\in\mathbb{R}^2,\xi_i} R^2 + \lambda \sum_{i=1}^n \xi_i & \\ \text{s.c.} & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i & \forall i = 1,\ldots n \\ & \xi_i \geq 0 & \forall\, i = 1, \cdots, n \end{array}$$

In this formulation the ξ_i represent slack variables (that is some samples can lie outside the sphere), $\lambda > 0$ is a regularization parameter (fixed by the user) which controls the trade-off between volume minimization (the radius) and the modeling errors. Higher values of λ will induce enclosing almost all the samples while small values may allow more samples to be outside the sphere.

1. Give the related lagrangian $\mathcal{L}$of this problem.
2. Write down the KKT stationary optimality conditions with relation to the primal variables $R, \mathbf{c}, \xi_k$. Deduce then the expression of the center $\mathbf{c}$ as a function of the training samples $\mathbf{x}_i$ and the Lagrange parameters.

3. Formulate the dual problem and propose a way to solve it.
4. By carefully justifying your answers, assess if the following samples may be support vectors.
   (a) samples strictly lying in the sphere
   (b) samples located on boundary of the sphere,
   (c) samples lying outside the sphere.
5. Using the KKT complementary conditions related to samples on the sphere, propose a way to estimate the radius R.
6. Let assume we learn from the training samples $\mathbf{c}$ and R. Let $\mathbf{z}$ any new sample. With the learned parameters how can we check if $\mathbf{z}$ is a "normal sample" or an outlier ?