Introduction à l'apprentissage automatique - TD3

Caio Corro

1 Hinge loss

The gold of this exercise is to show the equivalence between margin separation and the hinge loss.

1. **Binary classification.** We assume the output set is $Y = \{-1, 1\}$ and the training data $D = \left\{ (\boldsymbol{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{n}$. We consider the following training problem:

$$\begin{aligned} \min_{\boldsymbol{a}, b} \quad & \frac{\beta}{2} \|\boldsymbol{a}\|^2 \\ \text{s.t.} \quad & y^{(i)} \left(\boldsymbol{a}^\top \boldsymbol{x}^{(i)} + b \right) \geq m \qquad \forall 1 \leq i \leq n \end{aligned}$$

where $m > 0$ is a prefixed margin.

   (a) Give an interpretation of this mathematical program.
   (b) What issue can happen for a given dataset D?
   (c) How can we fix this problem? (think about allowing errors, but at the same time minimizing the number of errors)
   (d) Show that solving the problem from previous question is equivalent to using the hinge loss to train a linear model.

2. **Multiclass classification.** We assume the output set is $Y = E(k)$ and the training data $D = \left\{ (\boldsymbol{x}^{(i)}, y^{(i)}) \right\}_{i=1}^{n}$. We consider the following training problem:

$$\begin{aligned} \min_{\boldsymbol{A}, \boldsymbol{b}} \quad & \frac{\beta}{2} \|\boldsymbol{A}\|^2 \\ \text{s.t.} \quad & \langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}^{(i)} + \boldsymbol{b} \rangle + m \leq \langle \boldsymbol{y}^{(i)}, \boldsymbol{A}\boldsymbol{x}^{(i)} + \boldsymbol{b} \rangle \qquad \forall 1 \leq i \leq n, \boldsymbol{y} \in E(k) \setminus \{\boldsymbol{y}^{(i)}\} \end{aligned}$$

where $m > 0$ is a prefixed margin.

   (a) Give an interpretation of this mathematical program.
   (b) How many constraints are there?
   (c) Show that we can rewrite the problem with only m constraints.
   (d) Use the same trick as in the binary case to show that solving this problem is equivalent to using the hinge loss to train a linear model.

2 Gradient descent

1. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable function. The function $h_{\boldsymbol{a}} : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as:

$$h_{\boldsymbol{a}}(\boldsymbol{x}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{a} - \boldsymbol{x} \rangle$$

is a linear approximation of f around $\boldsymbol{a}$. Moreover, if f is convex, then $h_{\boldsymbol{a}}$ is a linear sub-estimator of f. Assume we want to approximately minimize the function f, i.e. solve:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}).$$

As this problem may be difficult, we may want to approximate it using an easier problem.
Consider the following problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} h_{\boldsymbol{a}}(\boldsymbol{x})$$

for a given $\boldsymbol{a} \in \mathbb{R}^d$. Why is solving this surrogate problem is useless?

2. Consider the following problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} h_{\boldsymbol{a}}(\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{a}\|^2$$

where $L > 0$ is a given constant. How does the term $\frac{L}{2}\|\boldsymbol{x} - \boldsymbol{a}\|^2$ impact the solution?

3. Show that the previous problem has a closed form solution. How can you interpret this solution?

4. Minimizing a function $f : \mathbb{R}^d \to \mathbb{R}$ using standard gradient descent method is simply computing a sequence of arguments as follows:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \epsilon \nabla f(\boldsymbol{\theta}^{(t)})$$

where $\epsilon > 0$ is the stepsize. A popular technique to avoid overfitting is weight decay, that is the gradient descent updates are replaced with the following equation:

$$\boldsymbol{\theta}^{(t+1)} = (1 - \lambda)\boldsymbol{\theta}^{(t)} - \epsilon \nabla f(\boldsymbol{\theta}^{(t)})$$

where $\lambda > 0$ is the weight decay parameter. Prove that, for gradient descent, weight decay is equivalent to adding a L2-regularization term to the objective.