Introduction to machine learning

Masters M2MO & MIDS

Stéphane Gaïffas

Université de Paris

Today

- Again binary classification
- The linear SVM
- Construction of the hinge loss

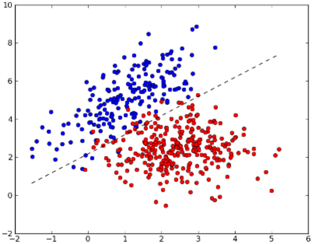
Setting

- Binary classification problem
- We observe a training dataset D of pairs (x_i, y_i) for $i = 1, \ldots, n$
- Features $x_i \in \mathbb{R}^d$ and labels $y_i \in \{-1, 1\}$
- Aim is to learn a classification rule that **generalizes** well
- Given a features vector $x \in \mathbb{R}^d$, we want to predict the label y
- Without **overfitting**

Linear classification. Why?

- Let's start simple!
- On very large datasets (n is large, say $n \geq 10^7$), no other choice (training complexity)
- Big data paradigm: lots of data $\Rightarrow$ simple methods are enough

A linear classifier



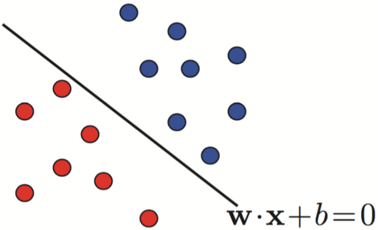
Learn $\hat{w} \in \mathbb{R}^d$ and $\hat{b}$ such that

$$\hat{y} = \mathrm{sign}(\langle x, \hat{w} \rangle + \hat{b})$$

is a good classifier

A dataset is **linearly separable** if we can find an hyperplane H that puts

- Points $x_i \in \mathbb{R}^d$ such that $y_i = 1$ on one side of the hyperplane
- Points $x_i \in \mathbb{R}^d$ such that $y_i = -1$ on the other
- H do not pass through a point x_i



An hyperplane

$$H = \{x \in \mathbb{R}^d : \langle w, x \rangle + b = 0\}$$

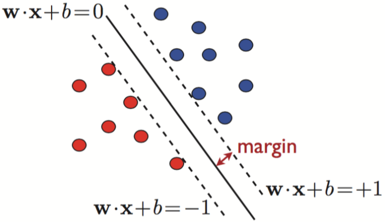
is a translation of a set of vectors orthogonal to w

- $w \in \mathbb{R}^d$ is a non-zero vector normal to the hyperplane
- $b \in \mathbb{R}$ is a scalar

Definition of H is invariant by multiplication of w and b by a non-zero scalar

If H do not pass through any sample point x_i, we can scale w and b so that

$$\min_{(x,y)\in D} |\langle w, x\rangle + b| = 1$$



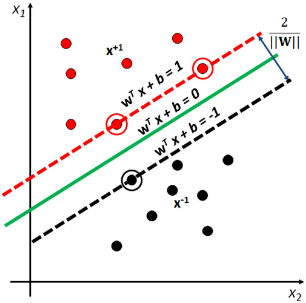
For such w and b, we call H the *canonical* hyperplane

The distance of any point $x' \in \mathbb{R}^d$ to H is given by

$$\frac{|\langle w, x' \rangle + b|}{\|w\|}$$

So, if H is a canonical hyperplane, its **margin** is given by

$$\min_{(x,y)\in D} \frac{|\langle w, x \rangle + b|}{\|w\|} = \frac{1}{\|w\|}.$$



In summary: if D is strictly linearly separable, we can find a canonical separating hyperplane

$$H = \{x \in \mathbb{R}^d : \langle w, x \rangle + b = 0\}.$$

that satisfies

$$|\langle w, x_i \rangle + b| \geq 1 \ \text{ for any } \ i = 1, \dots, n,$$

which entails that a point x_i is correctly classified if

$$y_i(\langle x_i, w \rangle + b) \geq 1.$$

The margin of H is equal to $1/\|w\|$.

Linear SVM: separable case

From that, we deduce that a way of classifying D with maximum margin is to solve the following problem:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2$$
$$\text{subject to} \quad y_i(\langle x_i, w\rangle + b) \geq 1 \ \text{ for all } \ i = 1, \ldots, n$$

Note that:

- This problem admits a **unique** solution
- It is a "quadratic programming" problem, which is easy to solve numerically
- Dedicated optimization algorithms can solve this on a large scale very efficiently

Some tools from **constrained optimization**

- Consider a constrained optimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x)$$
$$\text{subject to} \quad g_i(x) \le 0 \ \text{ for all } \ i = 1, \ldots, n$$

  where $f, g_1, \ldots, g_n : \mathbb{R}^d \to \mathbb{R}$
- We denote $P^* = f(x^*)$ the minimum of this objective (minimum of the **primal**)
- The associated **Lagrangian** is the function given on $\mathbb{R}^d \times \mathbb{R}_+^n$ by

$$L(x, \alpha) = f(x) + \sum_{i=1}^{n} \alpha_i g_i(x)$$

  for **Lagrange** or **dual** variables $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}_+^n$

- The **Lagrange dual** function is defined by

$$D(\alpha) = \inf_{x \in \mathbb{R}^d} L(x, \alpha) = \inf_{x \in \mathbb{R}^d} \Big(f(x) + \sum_{i=1}^{n} \alpha_i g_i(x) \Big)$$

  for $\alpha \in \mathbb{R}_+^n$
- D is always concave, as the infimum of linear functions
- We denote $D^* = D(\alpha^*) = \max_{\alpha \geq 0} D(\alpha)$ the optimal value of the dual. It is a convex problem (maximum of a concave function)
- For any **feasible** x and any $\alpha \geq 0$ we have $D(\alpha) \leq f(x)$, hence

$$D^* \leq P^*$$

  This is called the **weak duality** inequality and always holds
- Something that does not always holds is **strong duality**:

$$D^* = P^*$$

Strong duality holds under **constraint qualifications** (sufficient but not necessary)

Probably the best known one is **strong duality**:

- The primal problem is **convex**: $f, g_1, \ldots, g_n$ are convex
- **Slater**'s condition holds: there is some strictly feasible point $x \in \mathbb{R}^d$ such that

$$g_i(x) < 0 \quad \text{for all } i = 1, \ldots, n$$

- **Slater**'s condition is obvious for **affine** functions: inequality no longer strict, reduces to the original constraint $g_i(x) \leq 0$

Now, a fundamental tool: **KKT theorem** (Karush-Kuhn-Tucker)

- Assume that $f, g_1, \ldots, g_n$ are **differentiable**, assume **strong duality**.
- Then, $x^* \in \mathbb{R}^d$ is a solution of the primal problem if and only if there is $\alpha^* \in \mathbb{R}^n_+$ such that

$$
\begin{aligned}
\nabla_x L(x^*, \alpha^*) = \nabla f(x^*) + \sum_{i=1}^{n} \alpha_i^* \nabla g_i(x^*) = 0 \\
g_i(x^*) \le 0 \quad & \text{for any } i = 1, \ldots, n \\
\alpha_i^* g_i(x^*) = 0 \quad & \text{for any } i = 1, \ldots, n
\end{aligned}
$$

- These are known as the KKT conditions
- The last one is called **complementary slackness**

In summary: if

- primal problem is **convex** and
- constraint functions satisfy the **Slater**'s conditions

then

- **strong duality** holds.

If in addition we have that

- functions $f, g_1, \ldots, g_n$ are **differentiable**

then

- KKT conditions are **necessary and sufficient** for optimality

Back to the Linear SVM. The problem has the form

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} f(w)$$
$$\text{subject to} \quad g_i(w, b) \leq 0 \ \text{ for all } \ i = 1, \ldots, n$$

where

- $f(w) = \frac{1}{2}\|w\|_2^2$ is **strongly convex**, since

$$\nabla^2 f(w) = \boldsymbol{I}_d \succ 0$$

- Constraints are $g_i(w, b) \leq 0$ with **affine** functions

$$g_i(w, b) = 1 - y_i(\langle x_i, w \rangle + b)$$

  so that the constraints are **qualified**

We can apply the KKT theorem

Use this theorem to obtain a condition at the optimum

- It will lead to crucial properties on the SVM
- Allow to obtain the dual formulation of the problem

Lagragian

- Introduce dual variables $\alpha_i \geq 0$ for $i = 1, \ldots, n$ corresponding to the constraints $g_i(w, b) \leq 0$
- For $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $\alpha = (\alpha_1, \ldots \alpha_n) \in \mathbb{R}^n_+$, introduce the Lagrangian

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \alpha_i \big(1 - y_i(\langle w, x_i \rangle + b)\big)$$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \alpha_i \big(1 - y_i(\langle w, x_i\rangle + b)\big)$$

KKT conditions

Set the gradient to zero

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \text{namely} \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\nabla_b L(w,b,\alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{namely} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Write the complementary slackness condition

$$\alpha_i \big(1 - y_i(\langle w, x_i\rangle + b)\big) = 0 \quad \text{namely} \quad \alpha_i = 0 \ \text{ or } \ y_i(\langle w, x_i\rangle + b) = 1$$

for all $i = 1, \ldots, n$

This entails the following properties **at the optimum**

- There are **dual** variables $\alpha_i \geq 0$ such that the **primal** solution (w, b) satisfies

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

- We have that

$$\alpha_i \neq 0 \quad \text{iff} \quad y_i(\langle w, x_i \rangle + b) = 1$$

This means that

- w writes as a linear combination of the features vectors x_i that belong to the marginal hyperplanes $\{x \in \mathbb{R}^d : \langle w, x \rangle + b = \pm 1\}$
- These vectors x_i are called **support vectors**

The support vectors fully define the maximum-margin hyperplane, hence the name **Support Vector Machine**

Dual optimization problem

Lagrangian is

$$L(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \alpha_i \big(1 - y_i(\langle w, x_i\rangle + b)\big)$$

Plug $w = \sum_{i=1}^n \alpha_i y_i x_i$ in it to obtain

$$L(w,b,\alpha) = \frac{1}{2}\Big\| \sum_{i=1}^{n} \alpha_i y_i x_i \Big\|_2^2 + \sum_{i=1}^{n} \alpha_i - b \sum_{i=1}^{n} \alpha_i y_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

Recalling that $\sum_{i=1}^{n} \alpha_i y_i = 0$ and doing some algebra we arrive at the dual formulation

$$\max_{\alpha \in \mathbb{R}^n} \qquad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{subject to} \quad \alpha_i \geq 0 \;\; \text{and} \;\; \sum_{i=1}^{n} \alpha_i y_i = 0 \;\; \text{for all} \;\; i = 1, \ldots, n$$

Remarks

- As in the primal formulation, it is again a quadratic programming problem
- At optimum, we have (using KKT conditions) that the decision function is expressed using the dual variables as

$$x \mapsto \operatorname{sgn}\left(\langle w, x\rangle + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle x, x_i\rangle + b\right)$$

- The intercept b can be expressed for any support vector x_i as

$$b = y_i - \sum_{j=1}^{n} \alpha_j y_j \langle x_i, x_j\rangle$$

This allows to write the margin as a function of the dual variables

- Multiplying the last equality by $\alpha_i y_i$ and summing entails

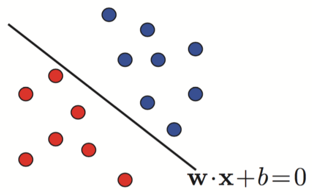
$$\sum_{i=1}^{n} \alpha_i y_i b = \sum_{i=1}^{n} \alpha_i y_i^2 - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

- Namely recalling that at optimum $\sum_{i=1}^n \alpha_i y_i = 0$ and $w = \sum_{i=1}^n \alpha_i y_i x_i$ we get

$$0 = \sum_{i=1}^{n} \alpha_i = \|w\|_2^2, \quad \text{namely}$$
$$\text{margin} = \frac{1}{\|w\|_2^2} = \frac{1}{\sum_{i=1}^n \alpha_i} = \frac{1}{\|\alpha\|_1}$$

- Okay, this is a nice theory, but...

Have you ever seen a dataset that looks that this?



Datasets are **not** linearly separable!

Keep cool and **relax** !

Replace the constraints

$$y_i(\langle w, x_i \rangle + b) \geq 1 \quad \text{for all} \quad i = 1, \ldots, n,$$

that are too strong, by the **relaxed** ones

$$y_i(\langle w, x_i \rangle + b) \geq 1 - s_i \quad \text{for all} \quad i = 1, \ldots, n,$$

for **slack variables** $s_1, \ldots, s_n \geq 0$

Slack rope

Linear SVM: non-separable case

Relax, but keep the slacks s_i as small as possible (goodness-of-fit)

Replace the original problem

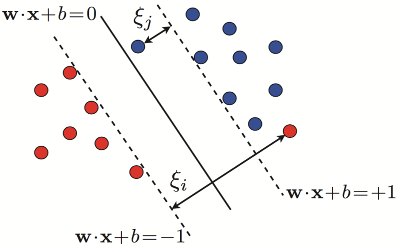
$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2$$
$$\text{subject to} \quad y_i(\langle x_i, w\rangle + b) \geq 1 \;\; \text{for all} \;\; i = 1, \ldots, n$$

by the relaxed one using slack variables:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, s \in \mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^n s_i$$
$$\text{subject to} \quad y_i(\langle x_i, w\rangle + b) \geq 1 - s_i \;\; \text{and} \;\; s_i \geq 0 \;\; \text{for all} \;\; i = 1, \ldots, n$$

where $C > 0$ is the "goodness-of-fit strength"

- The slack $s_i \geq 0$ measures the the distance by which x_i violates the desired inequality $y_i(\langle x_i, w\rangle + b) \geq 1$
- A vector x_i with $0 < y_i(\langle x_i, w\rangle + b) < 1$ is correctly classified but is an outlier, since $s_i > 0$
- If we omit outliers, training data is correctly classified by the hyperplane $\{x \in \mathbb{R}^d : \langle x, w\rangle + b = 0\}$ with a margin $1/\|w\|_2^2$
- The margin $1/\|w\|_2^2$ is called a **soft-margin** (in the non-separable case), while it is a **hard-margin** in the separable case



Linear SVM: non-separable case

So, we arrived at:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, s \in \mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} s_i$$

$$\text{subject to} \quad y_i(\langle x_i, w\rangle + b) \geq 1 - s_i \;\text{ and }\; s_i \geq 0 \;\text{ for all }\; i = 1, \ldots, n$$

Once again:

- This problem admits a **unique** solution
- It is a quadratic programming problem

The constant $C > 0$ is chosen using V-fold cross-valiation

Lagrangian

$$L(w, b, s, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} s_i$$
$$+ \sum_{i=1}^{n} \alpha_i \big(1 - s_i - y_i(\langle w, x_i\rangle + b)\big) - \sum_{i=1}^{n} \beta_i s_i$$

At optimum, let's again:

- set the gradients ∇_w, ∇_b and ∇_s to zero
- write the complementary conditions

$$\nabla_w L(w, b, s, \alpha, \beta) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \text{i.e.} \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\nabla_b L(w, b, s, \alpha, \beta) = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{i.e.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\nabla_s L(w, b, s, \alpha, \beta) = C - \alpha_i - \beta_i = 0 \quad \text{i.e.} \quad \alpha_i + \beta_i = C$$

and the complementary condition

$$\alpha_i\big(1 - s_i - y_i(\langle w, x_i\rangle + b)\big) = 0 \text{ i.e. } \alpha_i = 0 \ \text{ or } \ y_i(\langle w, x_i\rangle + b) = 1 - s_i$$

$$\beta_i s_i = 0 \quad \text{i.e.} \quad \beta_i = 0 \ \text{ or } \ s_i = 0$$

for all $i = 1, \ldots, n$

This means that

- $w = \sum_{i=1}^{n} \alpha_i y_i x_i$
- If $\alpha_i \neq 0$ we say that x_i is a support vector and in this case $y_i(\langle w, x_i \rangle + b) = 1 - s_i$
  - If $s_i = 0$ then x_i belongs to a margin hyperplane
  - If $s_i \neq 0$ then x_i is an outlier and $\beta_i = 0$ and then $\alpha_i = C$

Support vectors either belong to a marginal hyperplane, or are outliers with $\alpha_i = C$

Dual problem

- Plugging $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ in $L(w, b, s, \alpha, \beta)$ leads to the same formula as before

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

- with the constraints

$$\alpha_i \geq 0, \quad \beta_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i + \beta_i = C$$

that can be rewritten for as

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

for all $i = 1, \ldots, n$

Leading to the following **dual problem**

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \ \text{ and } \ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{ for all } \ i = 1, \ldots, n$$

- This is the same problem as before, but with the extra constraint

$$\alpha_i \leq C$$

- It is again a convex quadratic program

As in the linearly separable case, the label prediction is expressed using the dual variables as

$$x \mapsto \operatorname{sgn}\left(\langle w, x\rangle + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle x, x_i\rangle + b\right)$$

The intercept b can be expressed for a support vector x_i such that $0 < \alpha_i < C$ as

$$b = y_i - \sum_{j=1}^{n} \alpha_j y_j \langle x_i, x_j\rangle$$

A very important remark

The dual problem

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \;\text{ and }\; \sum_{i=1}^{n} \alpha_i y_i = 0 \;\text{ for all }\; i = 1, \ldots, n$$

and the label prediction (using dual variables)

$$x \mapsto \operatorname{sgn}\left(\langle w, x \rangle + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle x, x_i \rangle + b\right)$$

depends only on the features x_i via their **inner products** $\langle x_i, x_j \rangle$!

- This will be particularly important next week: **kernel methods**

The hinge loss

Going back to the primal problem

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, s\in\mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n s_i$$

subject to $\quad y_i(\langle x_i, w\rangle + b) \geq 1 - s_i \;$ and $\; s_i \geq 0 \;$ for all $\; i = 1, \dots, n$

We remark that it can be rewritten as

$$\operatorname*{argmin}_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \max\Big(0, 1 - y_i(\langle x_i, w\rangle + b)\Big).$$

Introducing the **hinge loss**

$$\ell(y, y') = \max(0, 1 - yy') = (1 - yy')_+,$$

the problem can be written as

$$\underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b).$$

Leads to an alternative understanding of the linear SVM.

Another natural loss is the 0/1 loss given by

$$\ell_{0/1}(y, z) = \mathbf{1}_{yz \leq 0}.$$

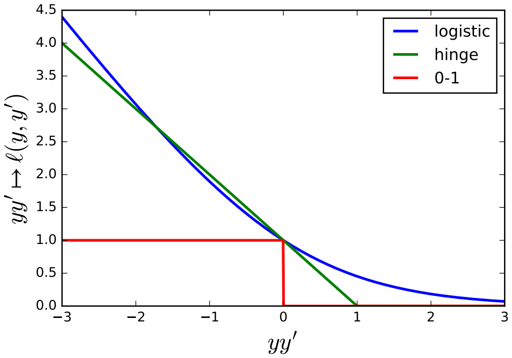
Instead of the Linear SVM, it would be nice to consider

$$\operatorname*{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \mathbf{1}_{y_i(\langle x_i, w\rangle + b) \leq 0},$$

but impossible numerically (NP-hard)

Hinge loss is a **convex surrogate** for the 0/1 loss

The losses we've seen so far for classification



$$\ell_{0-1}(y, y') = \mathbf{1}_{yy' \leq 0} \quad \ell_{\text{hinge}}(y, y') = (1 - yy')_+$$

$$\ell_{\texttt{logistic}}(y, y') = \log(1 + e^{-yy'}).$$

Grandmother's recipe:

Grandmother's recipes for logistic regression vs linear SVM

Logistic regression

- Logistic regression has a nice probabilistic interpretation
- Relies on the choice of the logit link function

SVM

- No model, only aims at separating points

No one is not better than the other in general. Depends on the data.

Once again, what is always important though is the **construction of the features** you'll use for training

Thank you!