Machine learning

Claire Boyer

September 23, 2020

Today

1. Convex losses

2. Linear Support Vector Machine (SVM)

Kernels:

- Mohri et al. "Foundations of machine learning"
- Thanks to Stéphane Gaïffas & Maxime Sangnier

More references: to come at the end of each section

Summary

Convexification

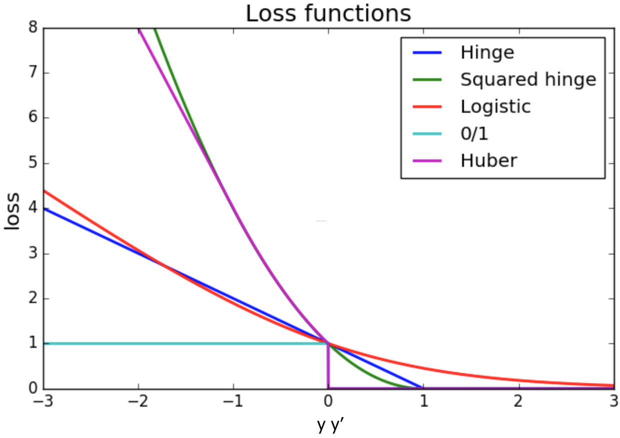
- The classification loss $\mathbb{1}_{g(x)\neq y}$ can be difficult to optimize (NP-hard).
- The idea is to smooth the indicator $(g(x), y) \mapsto \mathbb{1}_{g(x)\neq y}$.
- In the case of binary classification between -1 and 1, one has

$$\mathbb{E}\mathbb{1}_{Yf(X)<0} \leqslant \mathcal{R}(f) \leqslant \mathbb{E}\mathbb{1}_{Yf(X)\leqslant 0}$$

so one can rewrite the indicator as follows
$(g(x), y) \mapsto \mathbb{1}_{g(x)y\leqslant 0}$ Then we want to bound it from above with a convex function of $yg(x)(= yy')$.

Typical choices for convex loss functions

- Hinge loss (SVM) : $\ell(y, y') = (1 - yy')_+$
- Quadratic Hinge loss : $\ell(y, y') = \frac{1}{2}(1 - yy')_+^2$
- Huber loss : $\ell(y, y') = -4yy'\mathbb{1}_{yy'<-1} + (1 - yy')_+^2\mathbb{1}_{yy'\geqslant -1}$
- Logit loss : $\ell(y, y') = \log(1 + e^{-yy'})$



Question

What do we loose by convexifying the risk?

Question

What do we loose by convexifying the risk?

- Let us consider $\varphi : \mathbb{R} \to \mathbb{R}_+$ a loss function such that φ is strictly decreasing strictly convex, differentiable,

$$\varphi(0) = 1 \qquad \lim_{x \to \infty} \varphi(x) = 0.$$

Question

What do we loose by convexifying the risk?

- Let us consider $\varphi : \mathbb{R} \to \mathbb{R}_+$ a loss function such that φ is strictly decreasing strictly convex, differentiable,

$$\varphi(0) = 1 \qquad \lim_{x \to \infty} \varphi(x) = 0.$$

- One can then define the associated risk and its empirical version:

$$A(f) = \mathbb{E}[\varphi(Yf(X))] \quad \text{and} \quad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i f(X_i)).$$

About the "true" risk minimizer

Question

What is $f^\star = \operatorname{argmin}_f A(f)$ for φ strictly convex and differentiable?

About the "true" risk minimizer

Question

What is $f^\star = \text{argmin}_f A(f)$ for φ strictly convex and differentiable?

- Clearly, since $Y \in \{-1, 1\}$,

  $$\mathbb{E}[\varphi(Yf(X))|X = x] = r(x)\varphi(f(x)) + (1 - r(x))\varphi(-f(x)),$$

  with $r(x) = \mathbb{P}(Y = 1|X = x)$

About the "true" risk minimizer

Question

What is $f^\star = \operatorname{argmin}_f A(f)$ for φ strictly convex and differentiable?

- Clearly, since $Y \in \{-1, 1\}$,

$$\mathbb{E}[\varphi(Yf(X))|X = x] = r(x)\varphi(f(x)) + (1 - r(x))\varphi(-f(x)),$$

  with $r(x) = \mathbb{P}(Y = 1|X = x)$
- Consequence: $f^\star(x) = \operatorname{argmin}_\alpha h_{r(x)}(\alpha)$, where

$$h_r(\alpha) = r\varphi(\alpha) + (1 - r)\varphi(-\alpha), \quad r \in [0, 1].$$

About the "true" risk minimizer

Question

What is $f^\star = \operatorname{argmin}_f A(f)$ for φ strictly convex and differentiable?

- Clearly, since $Y \in \{-1, 1\}$,

$$\mathbb{E}[\varphi(Yf(X))|X = x] = r(x)\varphi(f(x)) + (1 - r(x))\varphi(-f(x)),$$

with $r(x) = \mathbb{P}(Y = 1|X = x)$

- Consequence: $f^\star(x) = \operatorname{argmin}_\alpha h_{r(x)}(\alpha)$, where

$$h_r(\alpha) = r\varphi(\alpha) + (1 - r)\varphi(-\alpha), \quad r \in [0, 1].$$

- Note: h_r is strictly convex and therefore $f^\star$ is well defined.
- The minimum is achieved for $h_r'(\alpha) = 0$, i.e.

$$\frac{r}{1 - r} = \frac{\varphi'(-\alpha)}{\varphi'(\alpha)}.$$

About the "true" risk minimizer

- Since φ' is strictly increasing, the solution is positive if and only if $r > 1/2$
- Conclusion: $f^\star(x) > 0$ iff $r(x) = \mathbb{P}(Y = 1|X = x) > 1/2$
- This is the Bayes classifier!

$$2\mathbb{1}_{f^\star(x)>0} - 1.$$

- Examples:
  - $\varphi(x) = e^{-x} \quad \Rightarrow \quad f^\star(x) = \frac{1}{2}\log(r(x)/(1 - r(x))$
  - $\varphi(x) = \text{hinge loss} \quad \Rightarrow \quad f^\star(x) = 2\mathbb{1}_{r(x)>0} - 1$. the Bayes classifier itself!

Connection between 0-1 function and φ-risk

Objective

- Connect $\mathcal{R}(f) - \mathcal{R}^\star$ with $A(f) - A^\star$.
- Tool: $H : [0, 1] \to \mathbb{R}$ defined by $H(r) = \inf_\alpha h_r(\alpha)$

Connection between 0-1 function and φ-risk

Objective

- Connect $\mathcal{R}(f) - \mathcal{R}^\star$ with $A(f) - A^\star$.
- Tool: $H : [0,1] \to \mathbb{R}$ defined by $H(r) = \inf_\alpha h_r(\alpha)$

Lemma

Let φ be a convex loss function such that the following hold:

(i) $f^\star > 0$ *iff* $r(x) > 1/2$

(ii) *There exist constants $c \geqslant 0$ and $s \geqslant 1$ satisfying*

$$\left|\frac{1}{2} - r\right|^s \leqslant c^s(1 - H(r)), \quad r \in [0,1].$$

Then, for any function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathcal{R}(f) - \mathcal{R}^\star \leqslant 2c(A(f) - A^\star)^{1/s}.$$

Connection between 0-1 function and φ-risk

H can be evaluated for different losses:

- Exponential: $H(r) = 2\sqrt{r(1-r)}$
- Logit: $H(r) = -r \log_2 r - (1-r)\log_2(1-r)$
- In both cases, $c = 1/\sqrt{2}$ and $s = 2$.
- Hinge: $H(r) = 2\min(r, 1-r)$, $\rightsquigarrow$ $c = 1/2$ and $s = 1$.

$$\mathcal{R}(f) - \mathcal{R}^\star \leqslant 2c(A(f) - A^\star)^{1/s}.$$

Excess risk of convex minimizers

Consider the class $\mathcal{C} =$ a class of ± 1base classifiers, and

$$\mathcal{F}_\lambda = \left\{ f = \sum_{j=1}^{N} c_j g_j : N \in \mathbb{N}, g_1, \ldots, g_N \in \mathcal{C}, \sum_{j=1}^{N} |c_j| = \lambda \right\}$$

Theorem

Let $f_n^\star \in \arg\min_{f \in \mathcal{F}_\lambda} A_n(f)$, using either the exponential or the logit loss function, and let $\delta \in (0,1)$. Then, with probability at least $1-\delta$,

$$\mathcal{R}(f_n^\star) - \mathcal{R}^\star \leqslant 2\left(8L_\varphi \lambda \sqrt{\frac{VC_{\mathcal{C}} \log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/2} + \sqrt{2}\left(\inf_{f \in \mathcal{F}_\lambda} A(f) - A^\star\right)^{1/2}.$$

Proof I

We have

$$
\begin{aligned}
\mathcal{R}(f_n^\star) - \mathcal{R}^\star &\leqslant \sqrt{2}(A(f_n^\star) - A^\star)^{1/2} \\
&\leqslant \sqrt{2}\big(A(f_n^\star) - \inf_{f \in \mathcal{F}_\lambda} A(f)\big)^{1/2} + \sqrt{2}\big(\inf_{f \in \mathscr{F}_\lambda} A(f) - A^\star\big)^{1/2} \\
&\leqslant 2\big(\sup_{f \in \mathcal{F}_\lambda} |A_n(f) - A(f)|\big)^{1/2} + \sqrt{2}\big(\inf_{f \in \mathcal{F}_\lambda} A(f) - A^\star\big)^{1/2} \\
&\leqslant 2\left(8L_\varphi \lambda \sqrt{\frac{VC_{\mathcal{C}} \log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/2} \\
&\qquad + \sqrt{2}\big(\inf_{f \in \mathcal{F}_\lambda} A(f) - A^\star\big)^{1/2}
\end{aligned}
$$

with probability at least δ. At the last step we used an upper bound for $\sup_{f \in \mathcal{F}_\lambda} |A_n(f) - A(f)|$ which is left to prove to the reader.

Excess risk of convex minimizers

$$\mathcal{R}(f_n^\star) - \mathcal{R}^\star \leqslant 2\left(8L_\varphi\lambda\sqrt{\frac{VC_{\mathcal{C}}\log(n+1)}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/2} + \sqrt{2}\left(\inf_{f\in\mathcal{F}_\lambda} A(f) - A^\star\right)^{1/2}.$$

Note that for

- Exponential: $L_\varphi = e^\lambda$ and $B = e^\lambda$.
- Logit: $L_\varphi = 1/\log(2)$ and $B = \log_2(1+e^\lambda)$.
- If $\inf A(f) - A^\star = 0$, then $\mathcal{R}(f) - \mathcal{R}^\star = O\left(\sqrt{\frac{\log(n)}{n}}\right)$
- The exponent in the rate is dimension-free!
- Convex optimization, nous voilà!
- And also boosting algorithms

Epilog on convexification

Take-home message

- By studying a convex surrogate risk, we control the approximation error
$$\mathcal{R}(f_n^\star) - \mathcal{R}^\star$$

- Statistics for high-dimensional data,
  by P. Bühlmann & S. Van de Geer
- Convexity,classification, and risk bounds,
  by P. Bartlett, M. Jordan, J. McAuliffe

Summary

Linear Support Vector Machine (SVM)

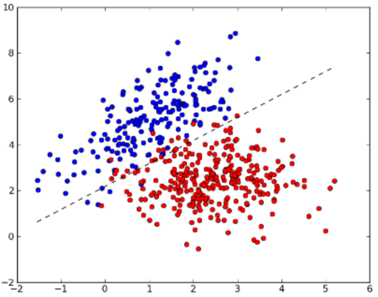
- Binary classification problem
- We observe a training dataset of pairs (x_i, y_i) for $i = 1, \ldots, n$
- Features $x_i \in \mathbb{R}^d$ and labels $y_i \in \{-1, 1\}$
- Aim is to learn a classification rule that **generalizes** well
- Given a features vector $x \in \mathbb{R}^d$, we want to predict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

Linear classification

Why?

- It's simple!
- On very large datasets (n is large, say $n \geqslant 10^7$), 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 $\hat{b}$ s.t.

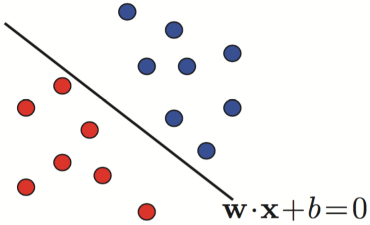
$$\hat{y} = \mathrm{sign}(\langle x, \hat{w} \rangle + \hat{b})$$

is a good classifier

Preliminary definitions

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



Preliminary definitions

An hyperplane

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

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

Canonical hyperplane

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

$$\min_{(x,y)\in D_n} |w^T x + b| = 1$$

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

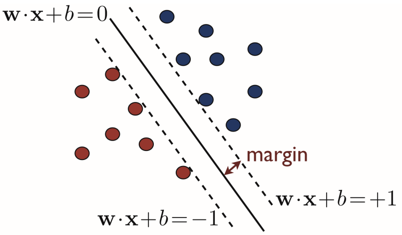


Figure: The marginal hyperplanes are the hyperplanes parallel to the separating hyperplane and passing through the closest points on the negative or positive sides.

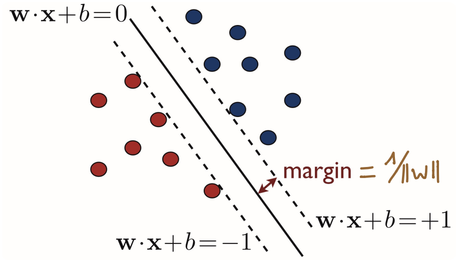
The margin

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

$$\min_{(x,y)\in D_n} \frac{|w^T x + b|}{\|w\|} = \frac{1}{\|w\|}.$$



Summary

If $\mathcal{D}_n$ is strictly linearly separable, we can find a canonical separating hyperplane

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

that satisfies

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

which entails that a point x_i is correctly classified if

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

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

Linear SVM: separable case

Maximum margin problem

A way of classifying $\mathcal{D}_n$ with maximum margin is to solve the following problem:

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

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

Tools from constrained convex optimization

- Consider a constrained optimization problem

$$
\begin{aligned}
\min_{x \in \mathbb{R}^d} \quad & f(x) \\
\text{subject to} \quad & h_i(x) = 0 \ \text{ for all } \ i = 1, \ldots, p \\
& g_j(x) \leqslant 0 \ \text{ for all } \ j = 1, \ldots, q
\end{aligned}
$$

  where $f, h_1, \ldots h_p, g_1, \ldots, g_q : \mathbb{R}^d \to \mathbb{R}$
- Denote $P^\star = f(x^\star)$ the minimum of the **primal** pb

Lagrangian

The associated **Lagrangian** is the function given on $\mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q_+$ by

$$
L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{q} \mu_j g_j(x)
$$

$\lambda = (\lambda_1, \ldots, \lambda_q) \in \mathbb{R}^p$, $\mu = (\mu_1, \ldots, \mu_q) \in \mathbb{R}^q_+$ are called **Lagrange** or **dual** variables.

The Lagrange dual function

$$
\begin{aligned}
D(\lambda, \mu) &:= \inf_{x \in \mathbb{R}^d} L(x, \lambda, \mu) \\
&= \inf_{x \in \mathbb{R}^d} \left(f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{q} \mu_j g_j(x) \right)
\end{aligned}
$$

for $(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}_+^q$

- D is always concave, as the infimum of linear functions
- Denote $D^\star := D(\lambda^\star, \mu^\star) = \max_{\substack{\lambda \\ \mu \geqslant 0}} D(\lambda, \mu)$ the optimal value of the dual. It is a convex problem (maximum of a concave function)

- For any **feasible** x and any $(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+$ we have $D(\lambda, \mu) \leqslant f(x)$, hence

Weak duality

$$D^\star \leqslant P^\star$$

This always holds!

- Something that does not always holds is

Strong duality

$$D^\star = P^\star$$

Lagrangian duality

Strong duality holds under

- convexity of the problem
- constraint qualifications

A simple way to have constraint qualification (sufficient but not necessary)

Slater's conditions

There is some strictly feasible point $x \in \mathbb{R}^d$ such that

$$\begin{aligned} h_i(x) &= 0 \quad \text{for all } i = 1, \dots, p \\ g_j(x) &< 0 \quad \text{for all } j = 1, \dots, q \end{aligned}$$

Karush-Kuhn-Tucker (KKT) conditions

(i) Assume that $f, g_1, \dots, g_q$ are **differentiable**, **convex**,

(ii) $h_1, \dots h_p$ are **affine** functions

(iii) Assume Slater's condition

NSC for optimality

Under (i), (ii), (iii),
$x^\star \in \mathbb{R}^d$ is a solution of the primal problem if and only if there is $(\lambda^\star, \mu^\star \in \mathbb{R}^p \times \mathbb{R}_+^q$ such that

$$\nabla_x L(x^\star, \lambda^\star, \mu^\star) = \nabla f(x^\star) + \sum_{i=1}^n \lambda_i^\star \nabla h_i(x^\star) + \sum_{j=1}^n \mu_j^\star \nabla g_j(x^\star) = 0$$

$$\begin{aligned} h_i(x^\star) &= 0 \quad \text{for any } i = 1, \dots, p \\ g_j(x^\star) &\leqslant 0 \quad \text{for any } j = 1, \dots, q \\ \mu_j^\star g_j(x^\star) &= 0 \quad \text{for any } j = 1, \dots, q \end{aligned}$$

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

Take-home message: Lagrangian duality

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

Exercise

Now that you know about the Lagrangian duality, you can prove that the distance of a point to the canonical hyperplane is well given by the formula on Slide 10!

Back to the linear SVM

The problem has the form

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

where

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

$$\nabla^2 f(w) = I_d \succ 0$$

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

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

  so that the constraints are **qualified**

Back to the linear SVM

KKT theorem

- Leads to crucial properties on the SVM
- Allows to obtain the dual formulation of the problem

Lagrangian

- Introduce dual variables $\mu_i \geqslant 0$ for $i = 1, \ldots, n$ corresponding to the constraints $g_i(w, b) \leqslant 0$
- For $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $\mu = (\mu_1, \ldots \mu_n) \in \mathbb{R}_+^n$, define the Lagrangian

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

KKT and linear SVM

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

KKT conditions

Set the gradient to zero

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

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

Write the complementary slackness condition: $\forall i = 1, \ldots, n$

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

SVM that's the name

At the optimum,

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

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

- We have that

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

SVM that's the name

At the optimum,

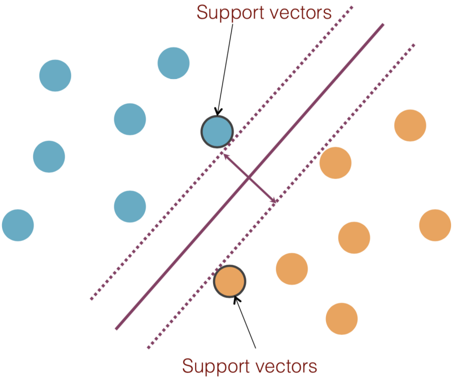
- There are **dual** variables $\mu_i \geqslant 0$ such that the **primal** solution (w, b) satisfies
$$w = \sum_{i=1}^{n} \mu_i y_i x_i$$
- We have that
$$\mu_i \neq 0 \quad \text{iff} \quad y_i(\langle w, x_i \rangle + b) = 1$$

This means that

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

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

SVM that's the name



Dual optimization problem I

Under strong duality, primal and dual problems are strongly related, and one can be used to solve the other.

- Recall that the Lagrangian is

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

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

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

Dual optimization problem II

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

Dual formulation

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

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

Comments

- 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

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

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

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

About the margin

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

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

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

About the margin

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

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

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

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

$$0 = \sum_{i=1}^{n} \mu_i = \|w\|_2^2 , \quad \text{namely}$$

$$\text{margin} = \frac{1}{\|w\|_2^2} = \frac{1}{\sum_{i=1}^{n} \mu_i} = \frac{1}{\|\mu\|_1}$$

About the margin

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

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

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

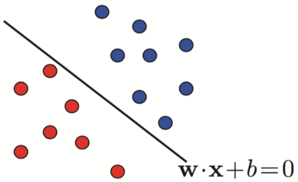
- Namely recalling that at optimum $\sum_{i=1}^n \mu_i y_i = 0$ and $w = \sum_{i=1}^n \mu_i y_i x_i$ we get

$$0 = \sum_{i=1}^n \mu_i = \|w\|_2^2\,, \qquad \text{namely}$$

$$\text{margin} = \frac{1}{\|w\|_2^2} = \frac{1}{\sum_{i=1}^n \mu_i} = \frac{1}{\|\mu\|_1}$$

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

Have you ever seen a dataset that looks like this?



Datasets are generally not linearly separable!

Relaxation for the non-separable case

Keep cool and **relax** !

Replace the constraints

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

Relaxation for the non-separable case

Keep cool and **relax** !

Replace the constraints

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

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

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

for slack variables $s_1, \ldots, s_n \geqslant 0$

Linear SVM: non-separable case

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

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

by the relaxed one using slack variables

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

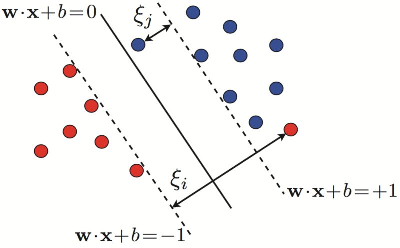
where $C > 0$ is the "goodness-of-fit strength"

- The slack $s_i \geqslant 0$ measures the the distance by which x_i violates the desired inequality $y_i(\langle x_i, w\rangle + b) \geqslant 1$

- The slack $s_i \geqslant 0$ measures the the distance by which x_i violates the desired inequality $y_i(\langle x_i, w\rangle + b) \geqslant 1$
- A vector x_i with $0 < y_i(\langle x_i, w\rangle + b) < 1$ is correctly classified but is an outlier, since $s_i > 0$

- The slack $s_i \geqslant 0$ measures the the distance by which x_i violates the desired inequality $y_i(\langle x_i, w\rangle + b) \geqslant 1$
- 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 $1/\left\|w\right\|_2^2$

- The slack $s_i \geqslant 0$ measures the the distance by which x_i violates the desired inequality $y_i(\langle x_i, w\rangle + b) \geqslant 1$
- 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 $1/\left\|w\right\|_2^2$
- The margin $1/\left\|w\right\|_2^2$ 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:

Relaxed margin problem

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

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

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

Linear SVM: non-separable case

Lagrangian

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

with $\mu_i \geqslant 0$ and $\beta_i \geqslant 0$.

At optimum, let's again:

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

Linear SVM: non-separable case

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

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

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

and the complementary condition

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

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

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

Linear SVM: non-separable case

This means that

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

Linear SVM: non-separable case

This means that

- $w = \sum_{i=1}^{n} \mu_i y_i x_i$
- If $\mu_i \neq 0$ 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 $\mu_i = C$

Linear SVM: non-separable case

This means that

- $w = \sum_{i=1}^{n} \mu_i y_i x_i$
- If $\mu_i \neq 0$ 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 $\mu_i = C$

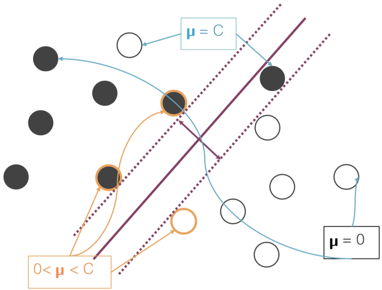
Support vectors either belong to a marginal hyperplane, or are outliers with $\mu_i = C$

Linear SVM: non-separable case

This means that

- $w = \sum_{i=1}^{n} \mu_i y_i x_i$
- If $\mu_i \neq 0$ 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 $\mu_i = C$

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



To the dual problem...

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

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

To the dual problem...

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

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

- with the constraints

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

that can be rewritten for as

$$0 \leqslant \mu_i \leqslant C, \quad \sum_{i=1}^n \mu_i y_i = 0$$

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

Dual problem

Dual problem

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

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

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

$$\mu_i \leqslant C$$

- It is again a convex quadratic program

Linear SVM: non-separable case

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

Labels given by

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

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

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

An important remark

The dual problem

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

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

and the label prediction (using dual variables)

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

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

- This will be particularly important later: **kernel methods**

SVM and the hinge loss

Going back to the primal problem

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

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

SVM and the hinge loss

Going back to the primal problem

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

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

We remark that it can be rewritten as follows.

Reformulation of the primal problem

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

SVM and the hinge loss

The hinge loss function

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

the problem can be written as

Reformulation of the primal problem

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

Leads to an alternative understanding of the linear SVM.

SVM and the hinge loss

Recall that the natural loss is the 0/1 one given by

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

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

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

but impossible numerically (NP-hard)

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

Conclusion

LDA/QDA

- Model: $X|Y \sim \mathcal{N}$

Logistic regression

- Logistic regression has a nice probabilistic interpretation
- Model $\mathrm{logit}(\mathbb{P}(Y = 1|X)$ is linear in X
- Relies on the choice of the logit link function
- ✗ does not work on separable dataset

SVM

- No model, only aims at separating points
- ✓ Thought for separable case
- ✓ But can be relaxed for the non-separable case