High dimensional statistics for genomic data

Laurent Jacob

January 30, 2017

The ridge regression

Ridge regression: bias and variance

- We observe n realizations of the previous linear model, represented by an $X \in \mathbb{R}^{n,p}$ matrix and a $Y \in \mathbb{R}^n$ vector.
- Consider the estimator

$$\hat{\theta} = \arg\min_{\theta} \left(\|Y - X\theta\|^2 + \lambda\|\theta\|^2 \right).$$

- We can show there exists a closed form for this estimator :

$$\hat{\theta} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

- **[Exercise]** Show that the bias $\mathbf{E}[\hat{\theta} - \bar{\theta}]$ of $\hat{\theta}$ is $-\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}$.

$$
\begin{aligned}
\mathbf{E}[\hat{\theta}] &= \mathbf{E}[(X^\top X + \lambda I)^{-1} X^\top Y] \\
&= \mathbf{E}[(X^\top X + \lambda I)^{-1} X^\top \left(X\bar{\theta} + \varepsilon\right)] \\
&= (X^\top X + \lambda I)^{-1} X^\top X \bar{\theta} + (X^\top X + \lambda I)^{-1} X^\top \mathbf{E}[\varepsilon] \\
&= (X^\top X + \lambda I)^{-1} X^\top X \bar{\theta}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}[\hat{\theta} - \bar{\theta}] &= (X^\top X + \lambda I)^{-1} X^\top X \bar{\theta} - \bar{\theta} \\
&= \left((X^\top X + \lambda I)^{-1} X^\top X - I\right) \bar{\theta} \\
&= (X^\top X + \lambda I)^{-1} (X^\top X - X^\top X - \lambda I) \bar{\theta} \\
&= -\lambda (X^\top X + \lambda I)^{-1} \bar{\theta}.
\end{aligned}
$$

Ridge regression: bias and variance

We now look at the variance of $\hat{\theta}$:

$$\begin{aligned} Var[\hat{\theta}] &= Var[(X^\top X + \lambda I)^{-1} X^\top Y] \\ &= (X^\top X + \lambda I)^{-1} X^\top Var[Y] X (X^\top X + \lambda I)^{-1} \\ &= \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1} \end{aligned}$$

(reminder : for a deterministic matrix A, $Var[AX] = AVar[X]A^\top$).

Ridge regression : bias and variance

Bias (1/3)

- The bias $-\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}$ increases with λ and tends to $-\bar{\theta}$.
- Remark: $\hat{\theta} \to 0$ when $\lambda \to \infty$, so the limit bias is the one incurred by estimating $\bar{\theta}$ by 0.
- If $\lambda = 0$ (unpenalized linear regression), the bias is zero.
- The amplitude of the bias also depends on the norm of $\bar{\theta}$: if the $\bar{\theta}$ which generated the data has a small norm, the bias/approximation error incurred by restricting ourselves to small norm estimators is smaller.

Bias (2/3)

- A little more precisely, the squared norm of the bias is ($\lambda \neq 0$) :

$$\begin{aligned} \| - \lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 &= \|(\lambda^{-1}X^\top X + I)^{-1}\bar{\theta}\|^2 \\ &= \|U\Sigma U^\top \bar{\theta}\|^2 = \|\Sigma U^\top \bar{\theta}\|^2, \end{aligned}$$

  where $U\Sigma U^\top$ is the spectral decomposition of $(\lambda^{-1}X^\top X + I)^{-1}$.
- The eigenvalues of $(\lambda^{-1}X^\top X + I)^{-1}$ are **[Exercise]** :

Bias (2/3)

- A little more precisely, the squared norm of the bias is $(\lambda \neq 0)$:

$$\begin{aligned} \| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 &= \|(\lambda^{-1}X^\top X + I)^{-1}\bar{\theta}\|^2 \\ &= \|U\Sigma U^\top \bar{\theta}\|^2 = \|\Sigma U^\top \bar{\theta}\|^2, \end{aligned}$$

  where $U\Sigma U^\top$ is the spectral decomposition of $(\lambda^{-1}X^\top X + I)^{-1}$.
- The eigenvalues of $(\lambda^{-1}X^\top X + I)^{-1}$ are **[Exercise]** :

$$\Sigma = \mathrm{Diag}\left(\lambda^{-1}e_i^2 + 1\right)^{-1} = \mathrm{Diag}\left(\frac{\lambda}{e_i^2 + \lambda}\right),$$

  where the e_i are the singular values of X.

Ridge regression: bias and variance

Bias (3/3)

$$\| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 = \left\| \mathrm{Diag}\left(\frac{\lambda}{e_i^2 + \lambda}\right) U^\top \bar{\theta} \right\|^2,$$

- Provides the shape of the convergence towards maximum bias as λ increases.
- If n/p is small, $X^\top X$ typically has small eigenvalues e_i^2, and the bias is larger (even more if θ is aligned with eigenvectors corresponding to small eigenvalues).
- Statistical interpretation: the bias is larger if the vector to be estimated lies in a direction of low empirical variance of the X.

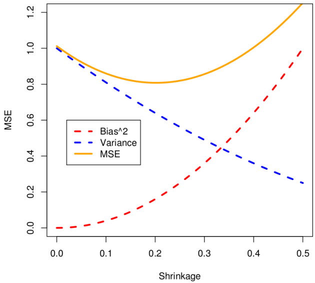
Variance

- Total variance

$$\begin{aligned}\mathrm{tr}\, Var[\hat{\theta}] &= \mathrm{tr}\left(\sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}\right)\\ &= \sigma^2 \mathrm{tr}\left((X^\top X + \lambda I)^{-2} X^\top X\right)\\ &= \sigma^2 \sum_i \frac{e_i^2}{(e_i^2 + \lambda)^2}.\end{aligned}$$

  tends to 0 as λ increases, and to the variance $\sigma^2 \mathrm{tr}(X^\top X)^{-1}$ of unpenalized linear regression as $\lambda \to 0$ (if $X^\top X$ is invertible).
- Here again if n/p is small, $X^\top X$ has small eigenvalues e_i^2 and the variance for $\lambda = 0$ increases.

Ridge regression: bias and variance



(from J. Taylor's slides)

- Illustrates the phenomenon we discussed abstractly on a particular estimator.
- In practice for a dataset, some λ tradeoffs yield smaller risks than others.
- λ can be chosen by hold out or cross validation (later in this class).

Ridge regression: remark

- We can also justify ridge regression from a numerical point of view: the λI term decreases the condition number of the $X^\top X$ matrix, which can otherwise get very small eigenvalues.
- A poorly conditionned $X^\top X$ leads to results which are very sensitive to small variations in the data.

Ridge regression: remark

- Historically, this motivated the introduction of ridge regression by Hoerl et Kennard (1970):

  > *We were charging \$90/day for our time, but had to charge \$450/hour for computer time [...], we found that we had both encountered the same phenomenon, one that had caused some embarrassment with clients. We found that multiple linear regression coefficients computed using least squares didn't always make sense when put into the context of the process generating the data. The coefficients tended to be too large in absolute value, some would even have the wrong sign, and they could be unstable with very small changes in the data.*

- Tikhonov (1943) and Philips (1962) already introduced Hilbert norm penalties to improve the conditionning of integral equation solutions.

Fundamentals of constrained optimization (from JP Vert)

Fundamentals of constrained optimization

Setting

- We consider an equality and inequality constrained optimization problem over a variable $x \in \mathcal{X}$:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h_i(x) = 0 \,, \quad i = 1, \ldots, m \,, \\ & g_j(x) \leq 0 \,, \quad j = 1, \ldots, r \,, \end{aligned}$$

  making **no assumption** of f, g and h.
- Let us denote by f^* the optimal value of the decision function under the constraints, i.e., $f^* = f\left(x^*\right)$ if the minimum is reached at a global minimum x^*.

Lagrangian and dual function

Lagrangian

The **Lagrangian** of this problem is the function $L : \mathcal{X} \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ defined by:

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

Lagrangian dual function

The **Lagrange dual function** $q : \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is:

$$\begin{aligned} q(\lambda, \mu) &= \inf_{x \in \mathcal{X}} L(x, \lambda, \mu) \\ &= \inf_{x \in \mathcal{X}} \left(f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) \right) . \end{aligned}$$

Properties of the dual function

- q **is concave in** (λ, μ), even if the original problem is not convex.
- The dual function yields lower bounds on the optimal value f^* of the original problem when μ is nonnegative:

$$q(\lambda, \mu) \leq f^* , \quad \forall \lambda \in \mathbb{R}^m, \forall \mu \in \mathbb{R}^r, \mu \geq 0 .$$

Proofs

- For each x, the function $(\lambda, \mu) \mapsto L(x, \lambda, \mu)$ is linear, and therefore both convex and concave in (λ, μ). The pointwise minimum of concave functions is concave, therefore q **is concave**.
- Let $\bar{x}$ be any feasible point, i.e., $h(\bar{x}) = 0$ and $g(\bar{x}) \leq 0$. Then we have, for any λ and $\mu \geq 0$:

$$\sum_{i=1}^{m} \lambda_i h_i(\bar{x}) + \sum_{i=1}^{r} \mu_i g_i(\bar{x}) \leq 0 \,,$$

$$\Longrightarrow \quad L(\bar{x}, \lambda, \mu) = f(\bar{x}) + \sum_{i=1}^{m} \lambda_i h_i(\bar{x}) + \sum_{i=1}^{r} \mu_i g_i(\bar{x}) \leq f(\bar{x}) \,,$$

$$\Longrightarrow \quad q(\lambda, \mu) = \inf_{x} L(x, \lambda, \mu) \leq L(\bar{x}, \lambda, \mu) \leq f(\bar{x}) \,, \quad \forall \bar{x} \,. \quad \square$$

Dual problem

Definition

For the (primal) problem:

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0\,, \quad g(x) \leq 0\,, \end{aligned}$$

the **Lagrange dual problem** is:

$$\begin{aligned} \textbf{maximize} \quad & q(\lambda, \mu) \\ \textbf{subject to} \quad & \mu \geq 0\,, \end{aligned}$$

where q is the (concave) Lagrange dual function and λ and μ are the Lagrange multipliers associated to the constraints $h(x) = 0$ and $g(x) \leq 0$.

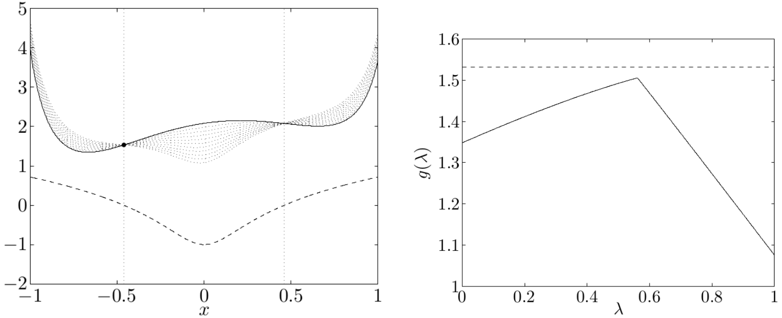
Weak duality

- Let d^* the optimal value of the Lagrange dual problem. Each $q(\lambda, \mu)$ is a lower bound for f^* and by definition d^* is the best lower bound that is obtained. The following **weak duality inequality** therefore **always holds**:

$$d^* \leq f^* \; .$$

- The difference $d^* - f^*$ is called the **optimal duality gap** of the original problem.

Illustration (from Boyd and Vandenberghe)



Strong duality

- We say that **strong duality** holds if the optimal duality gap is zero, i.e.:
$$d^* = f^* \,.$$
- If strong duality holds, then the best lower bound that can be obtained from the Lagrange dual function is **tight**.
- Strong duality does **not hold** for general nonlinear problems.
- It usually holds for **convex problems**.
- Conditions that ensure strong duality for convex problems are called **constraint qualification**.

Slater's constraint qualification

Strong duality holds for a **convex** problem:

$$
\begin{aligned}
\text{minimize} \quad & f(x) \\
\text{subject to} \quad & g_j(x) \leq 0 \,, \quad j = 1, \ldots, r \,, \\
& Ax = b \,,
\end{aligned}
$$

if it is **strictly feasible**, i.e., there exists at least one **feasible point** that satisfies:

$$
g_j(x) < 0 \,, \quad j = 1, \ldots, r \,, \quad Ax = b \,.
$$

Dual optimal pairs

Suppose that strong duality holds, x^* is primal optimal, (λ^*, μ^*) is dual optimal. Then we have:

$$
\begin{aligned}
f(x^*) &= q(\lambda^*, \mu^*) \\
&= \inf_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i^* h_i(x) + \sum_{j=1}^{r} \mu_j^* g_j(x) \right\} \\
&\leq f(x^*) + \sum_{i=1}^{m} \lambda_i^* h_i(x^*) + \sum_{j=1}^{r} \mu_j^* g_j(x^*) \\
&\leq f(x^*)
\end{aligned}
$$

Hence both inequalities are in fact **equalities**.

Complementary slackness

The second equality shows that:

$$\mu_j g_j(x^*) = 0 \ , \quad j = 1, \dots, r \ .$$

This property is called **complementary slackness**:
the ith optimal Lagrange multiplier is zero unless the ith constraint is active at the optimum.

Penalized vs constrained empirical risk minimization

Equivalence with a penalized estimator

In some cases, the constrained problem

$$\min_{\Omega(f)\leq\mu} \sum_{i=1}^{n} L(y_i, f(x_i)), \tag{1}$$

is equivalent in some sense to the penalized problem

$$\min_{f} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\Omega(f). \tag{2}$$

Any solution of (1) is a solution of (2) for some λ depending of μ, and vice-versa.

- The latter problem is sometimes easier to solve in practice.
- We will see later that estimators obtained by maximizing the **posterior likelihood** of some probabilistic models have this form.

Remark: equivalence with a penalized estimator

Example: L and Ω convex, $f \in \mathbb{R}^p$. We assume there exists f such that $\Omega(f) < \mu$. We note $L(f) \overset{\Delta}{=} \sum_{i=1}^n L(y_i, f(x_i))$ and

$$f_\mu \in \arg\min_{\Omega(f) \leq \mu} L(f),$$

$$f_\lambda \in \arg\min_f L(f) + \lambda \Omega(f).$$

Remark: equivalence with a penalized estimator

We first show that f_λ is a solution of the constrained problem for some μ **[Exercise]**:

Remark: equivalence with a penalized estimator

We first show that f_λ is a solution of the constrained problem for some μ **[Exercise]**:

- f_λ verifies the constraint of the constrained problem for $\mu = \Omega(f_\lambda)$.
- If there exists f' such that $L(f') < L(f_\lambda)$ and $\Omega(f') \leq \mu = \Omega(f_\lambda)$, then $L(f') + \lambda\Omega(f') < L(f_\lambda) + \lambda\Omega(f_\lambda)$ which contradicts the optimality of f_λ for the penalized problem.

Remark: equivalence with a penalized estimator

- This first part does not use convexity.
- We can therefore say in general that the regularization path of the penalized problem is included in the one of the constrained problem.

Remark: equivalence with a penalized estimator

We now show that f_μ is a solution of the penalized problem for some λ:

- Let $\mathcal{L}(f, \lambda) \stackrel{\Delta}{=} L(f) + \lambda(\Omega(f) - \mu)$ be the **Lagrangian** of the constrained problem (1).

Remark: equivalence with a penalized estimator

We now show that f_μ is a solution of the penalized problem for some λ:

- Let $\mathcal{L}(f, \lambda) \stackrel{\Delta}{=} L(f) + \lambda(\Omega(f) - \mu)$ be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is $q(\lambda) \stackrel{\Delta}{=} \min_f \mathcal{L}(f, \lambda)$.

Remark: equivalence with a penalized estimator

We now show that f_μ is a solution of the penalized problem for some λ:

- Let $\mathcal{L}(f, \lambda) \stackrel{\Delta}{=} L(f) + \lambda(\Omega(f) - \mu)$ be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is $q(\lambda) \stackrel{\Delta}{=} \min_f \mathcal{L}(f, \lambda)$.
- We note that here,

$$\min_f \mathcal{L}(f, \lambda) = \mathcal{L}(f_\lambda, \lambda). \tag{3}$$

Remark: equivalence with a penalized estimator

We now show that f_μ is a solution of the penalized problem for some λ:

- Let $\mathcal{L}(f, \lambda) \stackrel{\Delta}{=} L(f) + \lambda(\Omega(f) - \mu)$ be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is $q(\lambda) \stackrel{\Delta}{=} \min_f \mathcal{L}(f, \lambda)$.
- We note that here,

$$\min_f \mathcal{L}(f, \lambda) = \mathcal{L}(f_\lambda, \lambda). \tag{3}$$

- The dual always provides a lower bound to the primal solution: $\forall \lambda \geq 0, \min_{\Omega(f) \leq \mu} L(f) \geq q(\lambda) = \min_f \mathcal{L}(f, \lambda)$.

Remark: equivalence with a penalized estimator

We now show that f_μ is a solution of the penalized problem for some λ:

- Let $\mathcal{L}(f,\lambda) \stackrel{\Delta}{=} L(f) + \lambda(\Omega(f) - \mu)$ be the **Lagrangian** of the constrained problem (1).
- The **dual** of (1) is $q(\lambda) \stackrel{\Delta}{=} \min_f \mathcal{L}(f,\lambda)$.
- We note that here,

$$\min_f \mathcal{L}(f,\lambda) = \mathcal{L}(f_\lambda,\lambda). \tag{3}$$

- The dual always provides a lower bound to the primal solution: $\forall \lambda \geq 0$, $\min_{\Omega(f)\leq\mu} L(f) \geq q(\lambda) = \min_f \mathcal{L}(f,\lambda)$.
- Here by **strong duality** (obtained through Slater's conditions: convex problem and strictly feasible primal), we have

$$\min_{\Omega(f)\leq\mu} L(f) \stackrel{\text{s.d.}}{=} \max_{\lambda\geq 0} \min_f \mathcal{L}(f,\lambda) \stackrel{(3)}{=} \max_{\lambda\geq 0} \left(L(f_\lambda) + \lambda(\Omega(f_\lambda) - \mu)\right)$$

Remark: equivalence with a penalized estimator

$$\min_{\Omega(f)\leq\mu} L(f) = \max_{\lambda\geq 0}\left(L(f_\lambda) + \lambda(\Omega(f_\lambda) - \mu)\right)$$

- In addition, by Slater's conditions again, there exists λ^* such that

$$L(f_{\lambda^*}) + \lambda^*(\Omega(f_{\lambda^*}) - \mu) = \min_{\Omega(f)\leq\mu} L(f) = L(f_\mu).$$

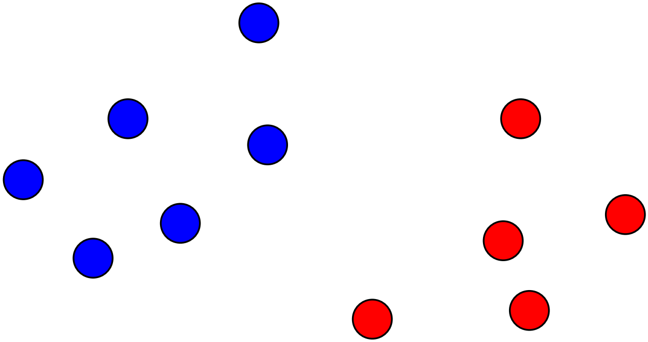
- By *complementary slackness*, it is necessary that $\lambda^*(\Omega(f_{\lambda^*}) - \mu) = 0$, which implies that $L(f_\mu) = L(f_\lambda^*)$ and:
  - either $\lambda^* = 0$ and $L(f_\mu) + 0\Omega(f_\mu) = L(f_\lambda^*) + 0\Omega(f_\lambda^*)$,
  - or $\Omega(f_{\lambda^*}) = \mu$ and $L(f_\mu) + \lambda^*\Omega(f_\mu) = L(f_\lambda^*) + \lambda^* \underbrace{\Omega(f_\mu)}_{\leq\mu=\Omega(f_{\lambda^*})} \leq L(f_\lambda^*) + \lambda^*\Omega(f_{\lambda^*})$.
- In both cases, f_μ is a solution of the problem penalized by λ^*.

Support vector machines

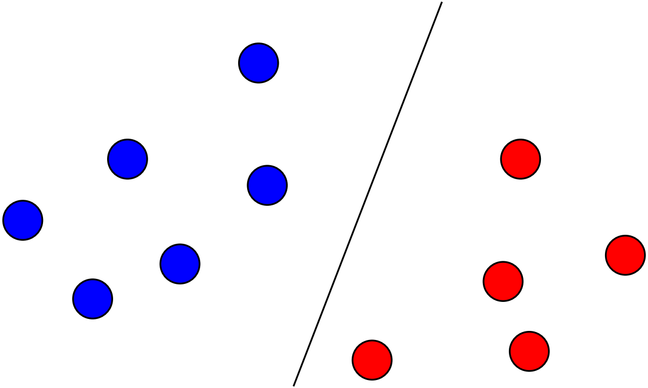
Linear SVM

- We now present Support Vector Machines, a classical statistical learning algorithm.
- Fits into the penalized/constrained empirical risk minimization framework.
- We choose the historical *large margin* presentation.

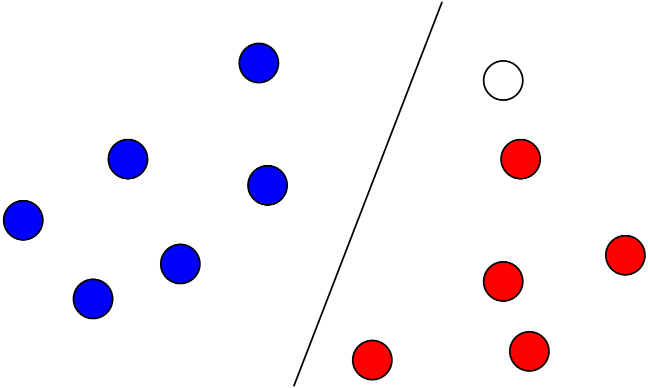
Linear classifier



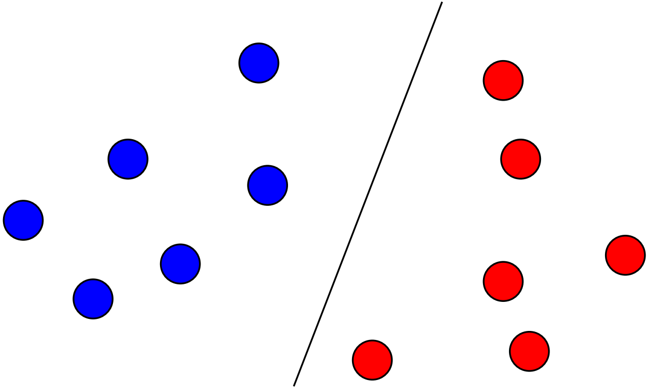
Linear classifier



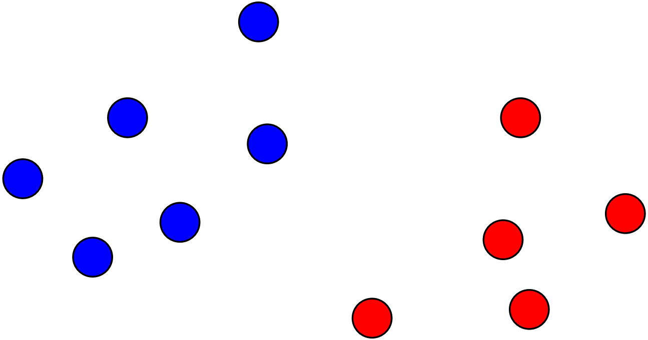
Linear classifier



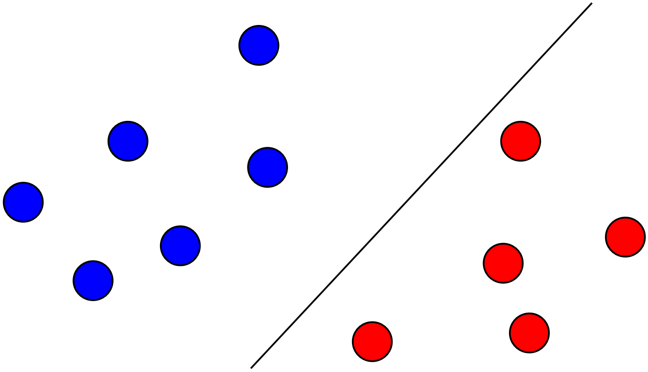
Linear classifier



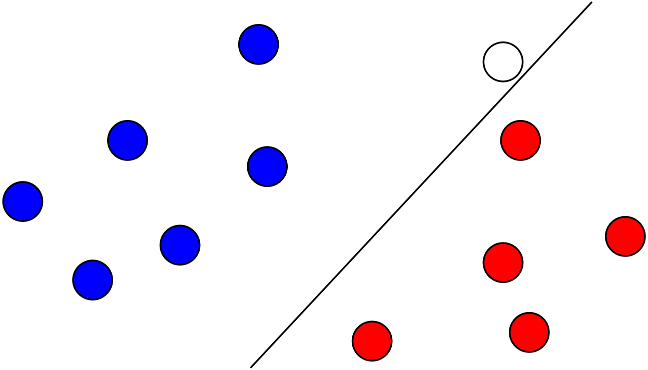
Linear classifier



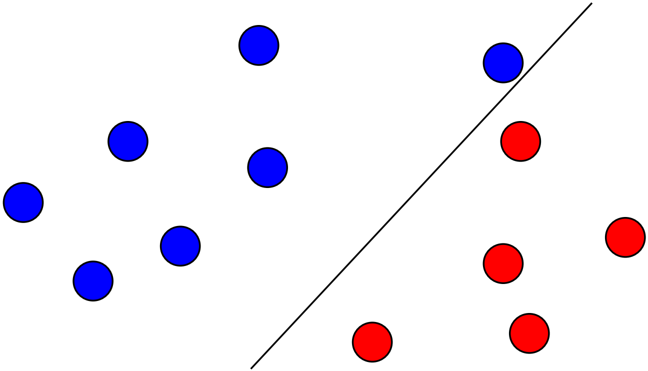
Linear classifier



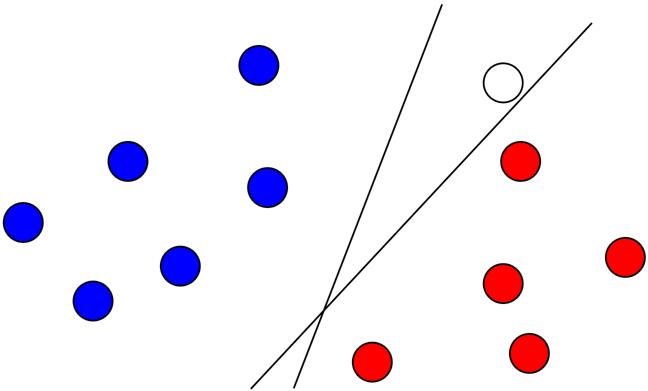
Linear classifier



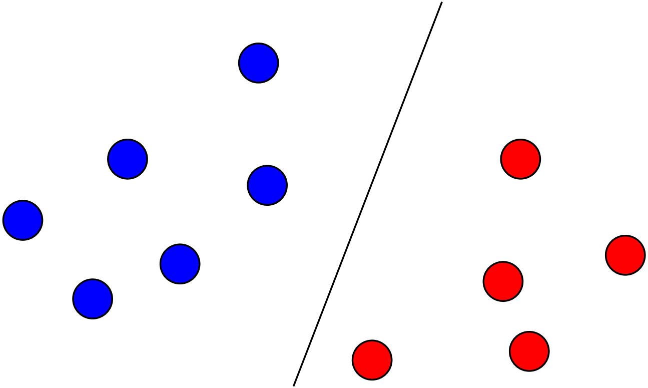
Linear classifier



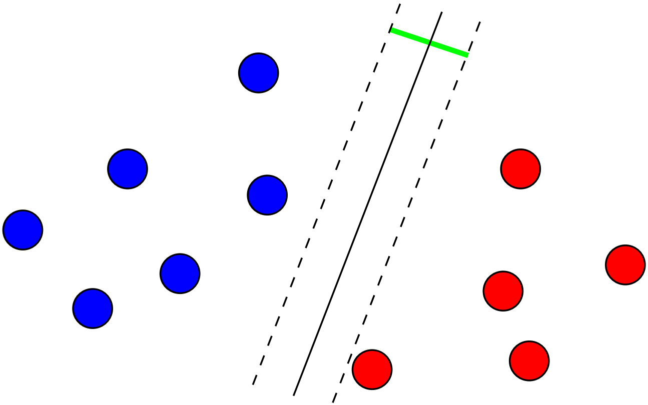
Which one is better?



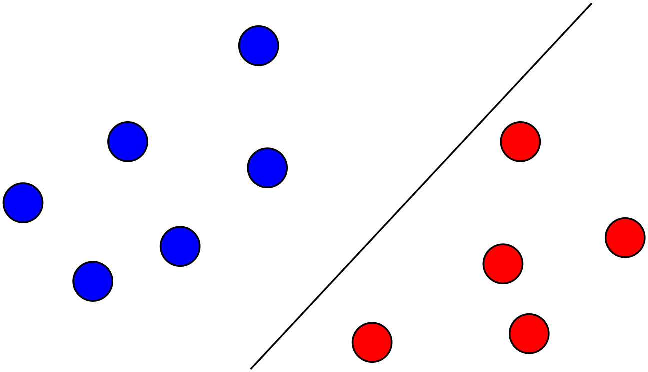
The margin of a linear classifier



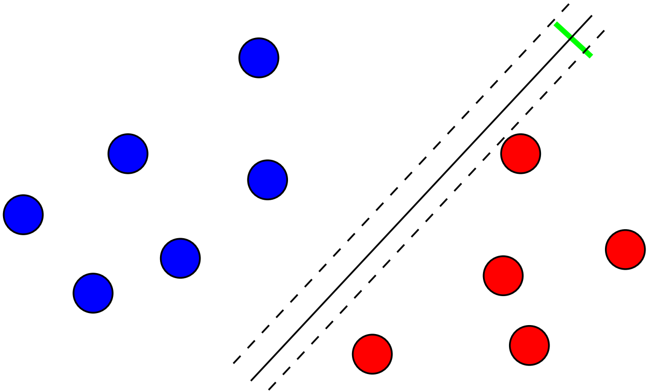
The margin of a linear classifier



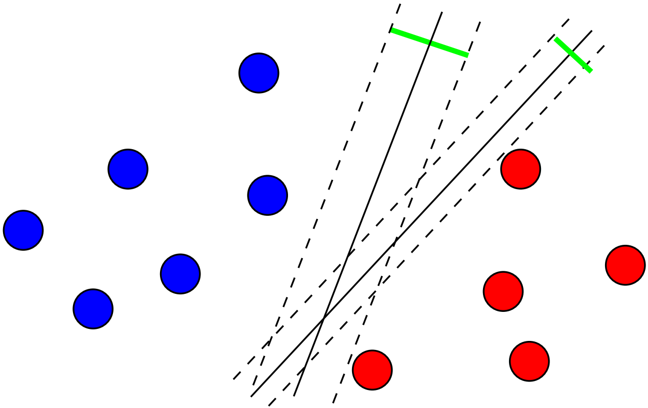
The margin of a linear classifier



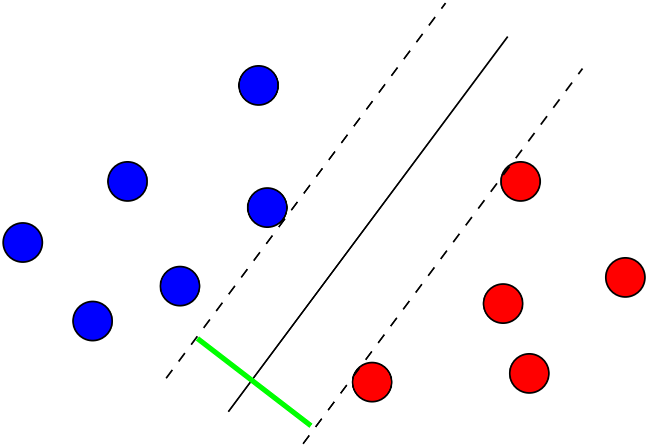
The margin of a linear classifier



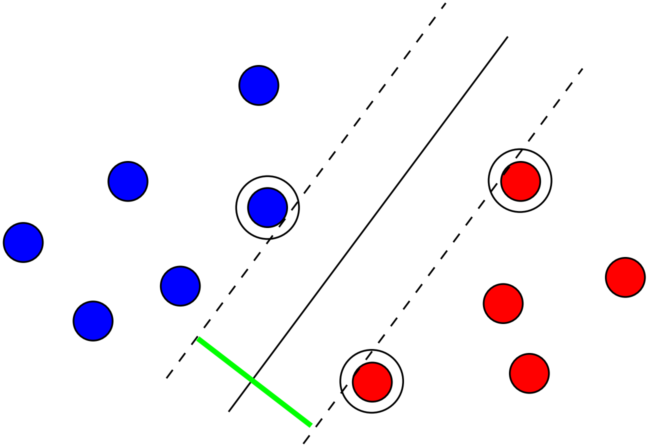
The margin of a linear classifier



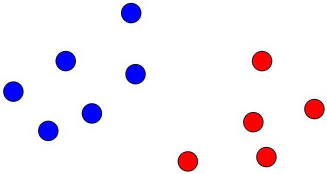
Largest margin classifier (*hard-margin SVM*)



Support vectors



More formally



- The **training set** is a finite set of n data/class pairs:

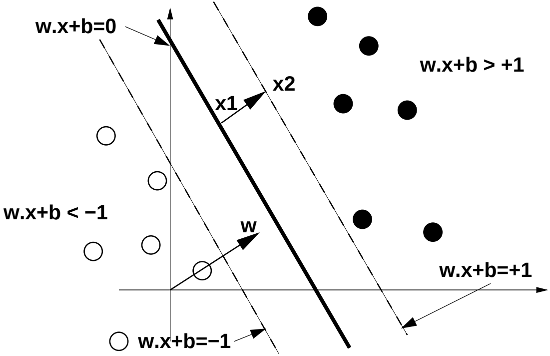
$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} ,$$

  where $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$.
- We assume (for the moment) that the data is **linearly separable**, i.e., that there exists $(w, b) \in \mathbb{R}^p \times \mathbb{R}$ such that:

$$\begin{cases} w.x_i + b > 0 & \text{if } y_i = 1 , \\ w.x_i + b < 0 & \text{if } y_i = -1 . \end{cases}$$

How to find the largest separating hyperplane?

For a given linear classifier $f(x) = w.x + b$ consider the "tube" defined by the values -1 and $+1$ of the decision function:



The margin is $2/\| w \|$

Indeed, the points x_1 and x_2 satisfy:

$$\begin{cases} w.x_1 + b = 0\,, \\ w.x_2 + b = 1\,. \end{cases}$$

By subtracting we get $w.(x_2 - x_1) = 1$, and therefore:

$$\gamma = 2\| x_2 - x_1 \| = \frac{2}{\| w \|}\,.$$

All training points should be on the correct side of the dotted line

For positive examples ($y_i = 1$) this means:

$$w.x_i + b \geq 1\,.$$

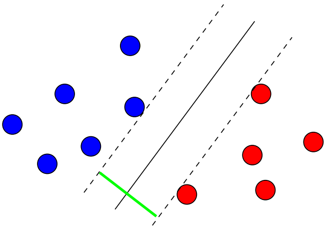
For negative examples ($y_i = -1$) this means:

$$w.x_i + b \leq -1\,.$$

Both cases are summarized by:

$$\forall i = 1, \ldots, n\,, \qquad y_i\left(w.x_i + b\right) \geq 1\,.$$

Finding the optimal hyperplane



Find (w, b) which minimize:

$$\| w \|^2$$

under the constraints:

$$\forall i = 1, \ldots, n \,, \qquad y_i \left(w.x_i + b \right) - 1 \geq 0 \,.$$

This is a classical quadratic program on $\mathbb{R}^{p+1}$.

Lagrangian

In order to minimize:

$$\frac{1}{2}\| w \|_2^2$$

under the constraints:

$$\forall i = 1, \ldots, n\,, \qquad y_i\left(w.x_i + b\right) - 1 \geq 0\,,$$

we introduce **one dual variable α_i for each constraint, i.e., for each training point**. The Lagrangian is:

$$L\left(w, b, \alpha\right) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i\left(y_i\left(w.x_i + b\right) - 1\right)\,.$$

Lagrangian

- $L(w, b, \alpha)$ is convex quadratic in w. It is minimized for:

$$\nabla_w L = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \Longrightarrow \quad \mathbf{w = \sum_{i=1}^{n} \alpha_i y_i x_i} \,.$$

- $L(w, b, \alpha)$ is affine in b. Its minimum is $-\infty$ except if:

$$\nabla_b L = \mathbf{\sum_{i=1}^{n} \alpha_i y_i = 0} \,.$$

Dual function

- We therefore obtain the **Lagrange dual function**:

$$q(\alpha) = \inf_{w \in \mathbb{R}^p, b \in \mathbb{R}} L(w, b, \alpha)$$
$$= \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j x_i . x_j & \text{if } \sum_{i=1}^n \alpha_i y_i = 0\,, \\ -\infty & \text{otherwise.} \end{cases}$$

- The dual problem is:

$$\begin{aligned} \text{maximize} \quad & q(\alpha) \\ \text{subject to} \quad & \alpha \geq 0\,. \end{aligned}$$

Dual problem

Find $\alpha^* \in \mathbb{R}^n$ **which maximizes**

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i . x_j,$$

under the (simple) constraints $\alpha_i \geq 0$ **(for** $i = 1, \ldots, n$**), and**

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

- This is a quadratic program on $\mathbb{R}^N$, with "box constraints". α^* can be found efficiently using dedicated optimization softwares.
- This dual shows how SVM are an instance of **kernel methods**.

Recovering the optimal hyperplane

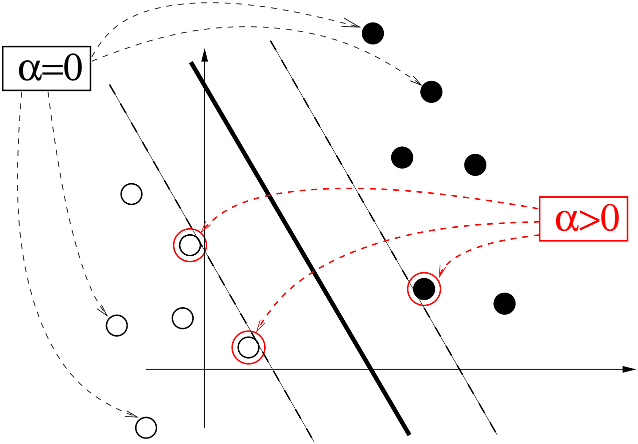
Once α^* is found, we recover (w^*, b^*) corresponding to the optimal hyperplane. w^* is given by:

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

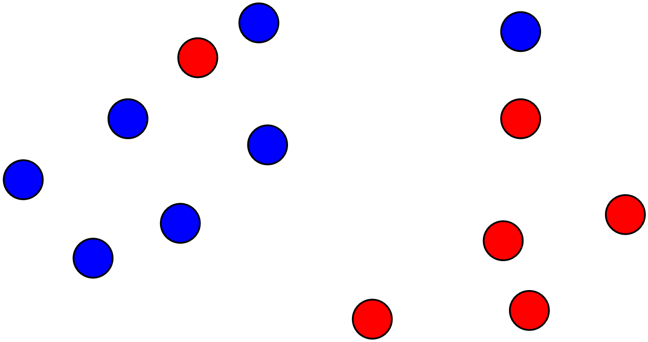
and the **decision function** is therefore:

$$\begin{aligned} f^*(x) &= w^*.x + b^* \\ &= \sum_{i=1}^{n} \alpha_i y_i x_i .x + b^* . \end{aligned} \tag{4}$$

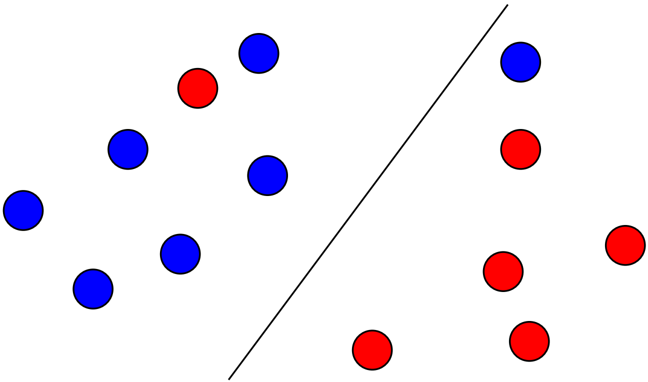
Interpretation: support vectors



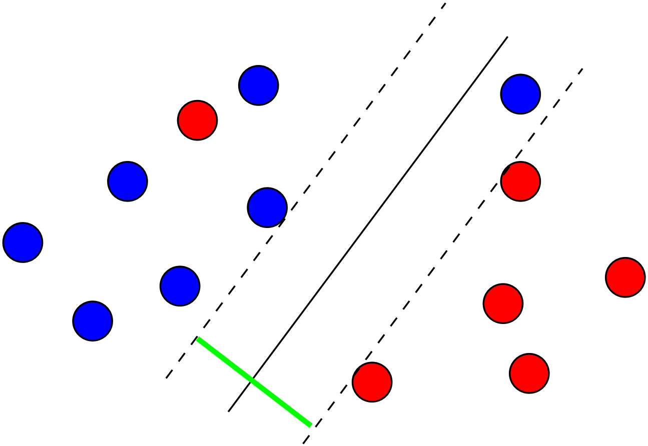
What if data are not linearly separable?



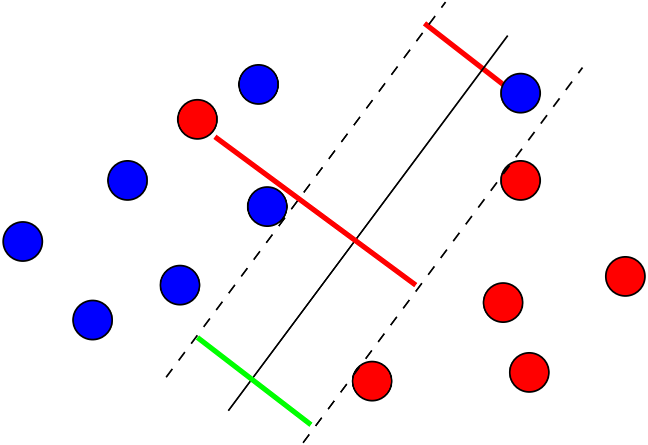
What if data are not linearly separable?



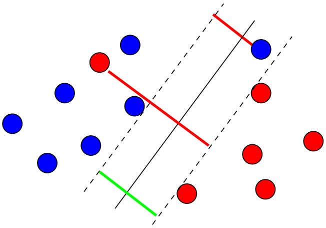
What if data are not linearly separable?



What if data are not linearly separable?



Relaxing the separation constraints

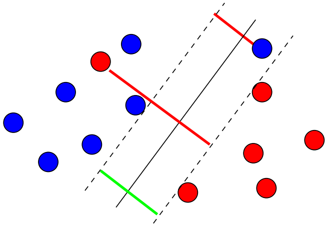


The problem is not feasible anymore. We need to **relax the separation constraints**:

$$\forall i = 1, \ldots, n\,, \qquad y_i\left(w.x_i + b\right) \geq 1 - \xi_i\,.$$

The ξ_i are called **slack variables**.

Trade-off



- Allowing a larger slack makes a larger margin possible.
- One way to control the trade-off is to integrate the slack variables as a cost in the objective function:

$$\begin{cases} \min_{w,b,\xi} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \forall i = 1, \ldots, n, \qquad y_i\,(w.x_i + b) \geq 1 - \xi_i \\ \forall i = 1, \ldots, n, \qquad \xi_i \geq 0 \end{cases}$$

- $C \in \mathbb{R}^+$ controls the trade-off.

Dual formulation of soft-margin SVM (*exercice*)

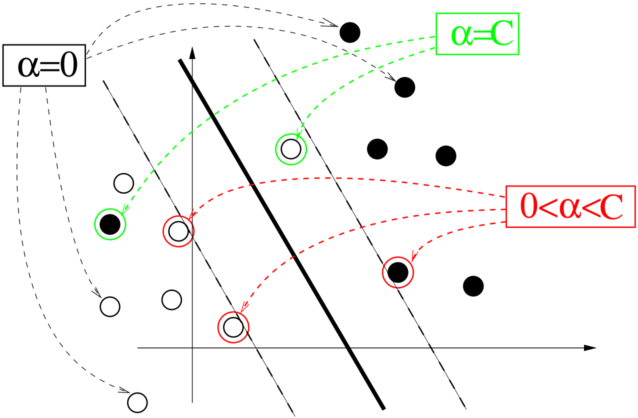
Maximize

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i . x_j,$$

under the constraints:

$$\begin{cases} 0 \leq \alpha_i \leq \mathbf{C}, & \text{for } i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases}$$

Interpretation: bounded and unbounded support vectors



Penalized empirical risk minimization

- Relaxed problem

$$\begin{cases} \min_{w,\xi} \|w\|^2 + C\sum_{i=1}^n \xi_i \\ \forall i = 1, \ldots, n\,, \qquad y_i\left(w.x_i + b\right) \geq 1 - \xi_i \\ \forall i = 1, \ldots, n\,, \qquad \xi_i \geq 0 \end{cases}$$

is equivalent to

$$\min_w \|w\|^2 + C\sum_{i=1}^n \max\left(0, 1 - y_i(w.x_i + b)\right)$$

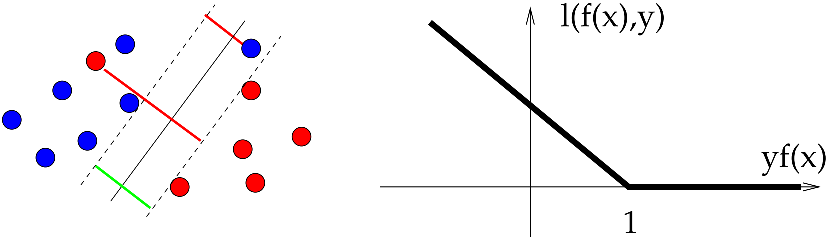
- Note: this is a useful trick to turn piecewise linear objective into linear objective with linear constraints.

Soft-margin SVM and hinge loss

$$\min_{w,b} \left\{ \sum_{i=1}^{n} \ell_{\text{hinge}} \left(w.x_i + b, y_i \right) + \lambda \| w \|_2^2 \right\} ,$$

for $\lambda = 1/C$ and the hinge loss function:

$$\ell_{\mathbf{hinge}}(u, y) = \max\left(1 - yu, 0\right) = \begin{cases} 0 & \textbf{if } yu \geq 1, \\ 1 - yu & \textbf{otherwise}. \end{cases}$$



Remarks

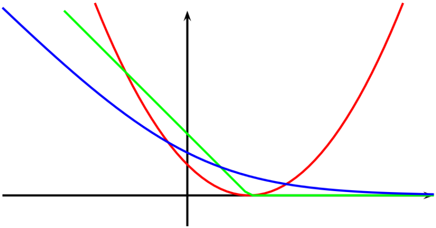
- We started from a different perspective (maximize margin) and showed retrospectively that the problem we solved could be thought of as penalized empirical risk minimization.
- Yields another interpretation for ℓ_2 regularization of linear functions.
- In practice controlling this trade-off makes sense even if the classes are linearly separable (as discussed during the first class).

Logistic regression (1/2)

A similar analysis can be made for logistic regression:

- Can be derived from a Bernouilli model:
- $\mathbf{E}[y_i|x_i] = p_i = \frac{1}{1+e^{-w^\top x_i}}$, where the logistic function ensures that $p_i \in [0,1]$.
- Leads to a linear separation: $\ln\left(\frac{p_i}{1-p_i}\right) = w^\top x_i$.
- Minimizing the negative log likelihood yields $\min_w \sum_{i=1}^n \ln(1 + e^{-w^\top x_i})$.

Logistic regression (2/2)



- Empirical risk for a loss function with very similar shape (and behavior) as the hinge loss.
- Intuition/justification is important but can be deceiving. In the end, it is crucial to compare objectives.

Algorithms

We mentioned hard and soft margin SVM could be written as a QP with box constraints. In practice however, faster dedicated algorithms were proposed, *e.g.*,

SimpleSVM

- Active set method: solve sub-problem with a restricted set of points, iteratively add the ones which most violate the constraints.
- Efficient when only a few α_i are non-zero (small C).

Stochastic gradient descent

- Take gradient steps with respect to randomly drawn single points.
- Efficient when the number of samples is large ("Large scale learning").

Penalties: non-exhaustive list of variations

- ℓ_1 norm (Vivian Viallon's class),
- Graph Laplacian,
- Trace norm,
- Fused norm,
- Group lasso,
- Other ℓ_p norms,
- Overlapping groups,
- Groups defined over a graph,
- Combinations
- ...

$\times$ combinations with various loss functions.