Machine Learning
Linear Classification

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 9, 22.11.2013

Linear Classification

Let $\mathcal{X} = \mathbb{R}^d$ be the input space, then the classifier $f : \mathbb{R}^d \to \{-1, 1\}$ has the form

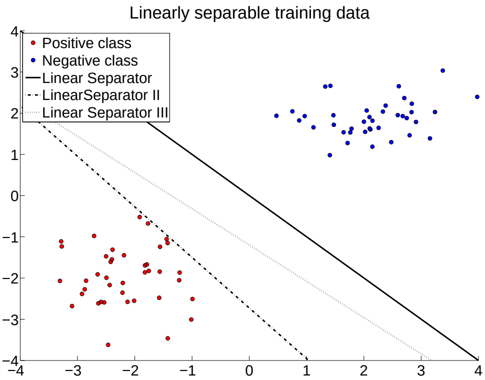
$$f(x) = \mathrm{sign}(\langle w, x\rangle + b) = \begin{cases} 1 & \text{if } \langle w, x\rangle + b > 0, \\ -1 & \text{if } \langle w, x\rangle + b \leq 0. \end{cases}$$

Separation of the input space $\mathbb{R}^d$ into two half spaces.

A training set $T = (X_i, Y_i)_{i=1}^n$ is **linearly separable** if there exists a weight vector w and an offset b such that,

$$Y_i\, f(X_i) = Y_i\,(\langle w, X_i\rangle + b) > 0, \qquad \forall i = 1, \ldots, n,$$

$\Rightarrow$ There exists a **hyperplane** $\{x \in \mathbb{R}^d \mid \langle w, x\rangle + b = 0\}$ which separates the sets $X_+ = \{X_i \in T \mid Y_i = 1\}$ and $X_- = \{X_i \in T \mid Y_i = -1\}$.



A training sample of a two-class problem in $\mathbb{R}^2$. The two classes are linearly separable and three different decision hyperplanes are shown which separate the two classes.

No distinction between the original input space $\mathcal{X} = \mathbb{R}^d$ and a possibly larger **feature space**, where we use basis functions/feature maps ϕ_i

$$x \in \mathbb{R}^d \longrightarrow (\phi_1(x), \ldots, \phi_D(x)),$$

to the feature space $\mathbb{R}^D$.

Functions are linear in the parameters but not necessarily linear in the input space !

Linear Classification III

No distinction between the original input space $\mathcal{X} = \mathbb{R}^d$ and a possibly larger **feature space**, where we use basis functions/feature maps ϕ_i

$$x \in \mathbb{R}^d \longrightarrow (\phi_1(x), \ldots, \phi_D(x)),$$

to the feature space $\mathbb{R}^D$.

Functions are linear in the parameters but not necessarily linear in the input space !

Definition

Let $g : \mathcal{X} \to \mathbb{R}$ be a function and $f(x) = \mathrm{sign}(g(x))$ be the resulting classifier with output in $\mathcal{Y} = \{-1, 1\}$, then we call the set

$$\{x \in \mathcal{X} \,|\, g(x) = 0\},$$

the **decision boundary** of the classifier f.

Three linear methods:

- **Linear Discriminant Analysis**,
- **Logistic Regression**,
- **Support Vector Machines**.

All three methods construct a **linear** classifier but all three have different **objectives**.

Properties and Motivation:

- often also called **Fisher Discriminant Analysis** named after its inventor Ronald A. Fisher, the "father" of parametric statistics.
- In **linear** classification the data x enters the classifier only via the inner product $\langle w, x \rangle$ with the weight vector.
  - Projection of the feature space $\mathbb{R}^D$ onto the line $L = \{\alpha w \,|\, \alpha \in \mathbb{R}\}$,
  - Classification of the data by thresholding.

What is the best projection in the sense that it optimally separates the data ?

Definitions:

- The class **centroids** m_+ and m_- of the positive and negative class are defined as:

$$m_+ = \frac{1}{n_+} \sum_{\{i \,|\, Y_i = 1\}} X_i, \qquad m_- = \frac{1}{n_-} \sum_{\{i \,|\, Y_i = -1\}} X_i,$$

  where $n_+ = |\{i \,|\, Y_i = 1\}|$ and $n_- = |\{i \,|\, Y_i = -1\}|$.

- The **within-class covariances** of the projections of the positive and negative class are given by

$$\sigma^2_{w,+} = \sum_{\{i \,|\, Y_i = 1\}} \Big(\langle w, X_i \rangle - \langle w, m_+ \rangle \Big)^2,$$

$$\sigma^2_{w,-} = \sum_{\{i \,|\, Y_i = -1\}} \Big(\langle w, X_i \rangle - \langle w, m_- \rangle \Big)^2.$$

LDA III

Criterion:

- - Large Distance of the projected class centroids $\langle w, m_+ \rangle$ and $\langle w, m_- \rangle$,
  - Small variances around the projected class centroids.
- The **Fisher criterion** is defined as

$$J(w) = \frac{\langle w, m_+ - m_- \rangle^2}{\sigma_{w,+}^2 + \sigma_{w,-}^2}.$$

Fisher criterion in matrix formulation:
The **between-class covariance** matrix C_B is defined as

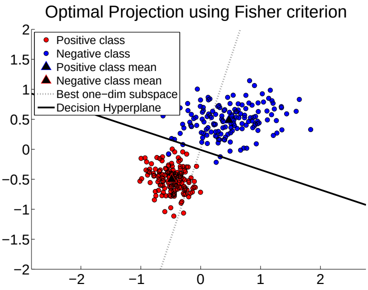
$$C_B = (m_+ - m_-)(m_+ - m_-)^T,$$

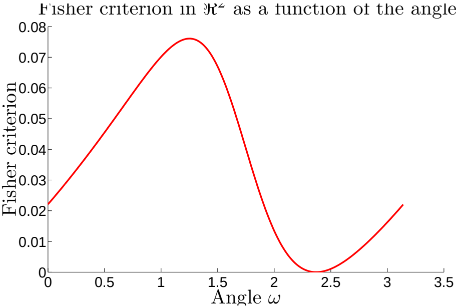
and the total **within-class covariance** matrix C_W as

$$C_W = \sum_{\{i \,|\, Y_i = 1\}} (X_i - m_+)(X_i - m_+)^T + \sum_{\{i \,|\, Y_i = -1\}} (X_i - m_-)(X_i - m_-)^T.$$

Then the **Fisher criterion** $J(w)$ can be written as

$$J(w) = \frac{\langle w, C_B w \rangle}{\langle w, C_W w \rangle}.$$





Left: Projection w optimizing the Fisher criterion and the optimal projection line $\{\alpha w + \frac{1}{2}(m_+ + m_-) \,|\, \alpha \in \mathbb{R}\}$. **Right:** The Fisher criterion as a function of the angle ω, where ω is a parameterization of all weight vectors $w = (\cos(\omega), \sin(\omega))$ in $\mathbb{R}^2$.

Optimal Projection

Lemma

The **optimal projection** $w^* = \arg\max_{w \in \mathbb{R}^D} J(w)$ *is given by*

$$w^* = C_W^{-1}(m_+ - m_-).$$

Proof: We have

$$\nabla_w J(w) = 2\frac{1}{\langle w, C_W w\rangle} C_B w - 2\frac{\langle w, C_B w\rangle}{\langle w, C_W w\rangle^2} C_W w.$$

We solve for the extrema of $J(w)$ and get

$$\frac{\langle w, C_W w\rangle}{\langle w, C_B w\rangle} C_B w = C_W w.$$

Now, $C_B w$ is always proportional to $m_+ - m_-$ and $\frac{\langle w, C_W w\rangle}{\langle w, C_B w\rangle}$ is just a scalar factor. Therefore

$$w^* \propto C_W^{-1}(m_+ - m_-).$$

LDA VI

- **Final classifier:**
$$f(x) = \operatorname{sign}(\langle w, x\rangle + b).$$
Determine **the threshold** b by minimizing the training error.
- Optimal Projection can also be derived using **least squares**. This yields the following optimization problem
$$(w', w_0') = \operatorname*{arg\,min}_{w\in\mathbb{R}^D, w_0\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - \langle w, X_i\rangle - w_0)^2.$$
One can prove (exercise)
$$w^* \sim w'.$$

Dimensionality Reduction

In **Dimensionality Reduction** we would like to have

- a lower dimensional $m \ll D$ representation of the data,
- which preserves the "interesting" properties of the data
  $\Rightarrow$ In classification: classifier should perform on the new m-dimensional space as well as on the original D-dimensional space.

The **between-class covariance** matrix C_B is defined as

$$C_B = \sum_{k=1}^{K} n_k (m_k - m)(m_k - m)^T.$$

and the total **within-class covariance** matrix C_W as

$$C_W = \sum_{k=1}^{K} \sum_{\{i \,|\, Y_i = k\}} (X_i - m_k)(X_i - m_k)^T,$$

The **Fisher criterion** $J(w)$ stays the same

$$J(w) = \frac{\langle w, C_B w \rangle}{\langle w, C_W w \rangle}.$$

One needs generally a $K-1$-dimensional subspace in order to separate K classes !

Rayleigh-Ritz principle

Proposition (Rayleigh-Ritz principle)

Let $A \in \mathbb{R}^{d\times d}$ *be a* **symmetric matrix**, *then*

$$\lambda_{\max} = \max_{x \in \mathbb{R}^d} \frac{\langle x, Ax\rangle}{\langle x, x\rangle},$$

is the largest eigenvalue of A *and the maximizing argument* $x_{\max}$ *is the corresponding eigenvector. Equivalently,*

$$\lambda_{\max} = \max_{x \in \mathbb{R}^d,\ \|x\|=1} \langle x, Ax\rangle\,.$$

Other eigenvalues and eigenvectors can be found as follows. Denote by $u_1, \ldots, u_r$ *the eigenvectors corresponding to the largest* r *eigenvalues, then the* $r+1$ *largest eigenvalue can be found as,*

$$\lambda_{r+1} = \max_{x \in \mathbb{R}^d,\ \langle x, u_s\rangle = 0,\ s=1,\ldots,r} \frac{\langle x, Ax\rangle}{\langle x, x\rangle}.$$

How can we get more projections from the Fisher criterion ?

$$J(w) = \frac{\langle w, C_B w\rangle}{\langle w, C_W w\rangle}.$$

The Fisher criterion can be seen as the variational formulation of the **generalized eigenvalue problem**

$$C_B w = \lambda\, C_W w.$$

Generalized Rayleigh-Ritz Principle!

m-dimensional projection is determined by the m eigenvectors corresponding to the m largest eigenvectors.

Logistic Regression:

Original Formulation: Maximum likelihood estimation using the following model for the conditional probability

$$\mathrm{P}(Y = 1 | X = x, w) = \frac{1}{1 + e^{-\langle w, \phi(x) \rangle}}.$$

Definition

Given a training sample $T_n = (X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$ and the function space $\mathcal{F} = \{\sum_{i=1}^D w_i \phi_i(x) \,|\, w \in \mathbb{R}^D\}$ we define **logistic regression** as the mapping $\mathcal{A} : T_n \to \mathcal{F}$ with,

$$T_n \mapsto f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \log\Big(1 + \exp(-Y_i \, \langle w, \phi(X_i) \rangle)\Big). \qquad (1)$$

Empirical risk minimization using the logistic loss !

- no analytical solution,
- solution using a Newton-type gradient descent method. Gradient and the Hessian of the empirical risk:

$$R_{\mathrm{emp}}(w) = \frac{1}{n}\sum_{i=1}^{n} \log\Big(1 + \exp(-Y_i \,\langle w, \phi(X_i)\rangle)\Big),$$

as

$$\frac{\partial R_{\mathrm{emp}}}{\partial w_s}(w) = -\frac{1}{n}\sum_{i=1}^{n} y_i\, \phi_s(X_i) \frac{\exp(-Y_i \,\langle w, \phi(X_i)\rangle)}{1 + \exp(-Y_i \,\langle w, \phi(X_i)\rangle)},$$

$$\frac{\partial^2 R_{\mathrm{emp}}}{\partial w_r \partial w_s}(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_s(X_i)\phi_r(X_i) \frac{\exp(-Y_i \,\langle w, \phi(X_i)\rangle)}{\Big(1 + \exp(-Y_i \,\langle w, \phi(X_i)\rangle)\Big)^2}.$$

Newton-Raphson algorithm: with stepsize fixed to 1,

$$w_{\mathrm{new}} = w_{\mathrm{old}} - \Big(\frac{\partial^2 R_{\mathrm{emp}}}{\partial w_r \partial w_s}(w)\Big)^{-1} \nabla_w R_{\mathrm{emp}}(w),$$

With the diagonal matrices W and D with diagonal entries

$$W_{ii} = \frac{\exp(-Y_i \langle w, \phi(X_i)\rangle)}{(1+\exp(-Y_i \langle w, \phi(X_i)\rangle))^2}, \qquad D_{ii} = \frac{\exp(-Y_i \langle w, \phi(X_i)\rangle)}{1+\exp(-Y_i \langle w, \phi(X_i)\rangle)},$$

we can write the gradient and Hessian $H(R_{\mathrm{emp}})$ of R_{emp} as

$$\nabla_w R_{\mathrm{emp}}(w) = -\frac{1}{n}\Phi^T D Y, \qquad H(R_{\mathrm{emp}})\Big|_w = \frac{1}{n}\Phi^T W \Phi.$$

Logistic Regression IV

Thus we can write the **Newton-Raphson update** as

$$\begin{aligned} w_{\mathrm{new}} &= w_{\mathrm{old}} + \left(\Phi^T\, W\, \Phi\right)^{-1} \Phi^T DY \\ &= \left(\Phi^T\, W\, \Phi\right)^{-1} \Phi^T W \left(\Phi w_{\mathrm{old}} + W^{-1} DY\right) = \left(\Phi^T\, W\, \Phi\right)^{-1} \Phi^T W Z, \end{aligned}$$

with $Z = \Phi w_{\mathrm{old}} + W^{-1} DY$.

Thus we can write the **Newton-Raphson update** as

$$\begin{aligned} w_{\mathrm{new}} &= w_{\mathrm{old}} + \left(\Phi^T W \Phi\right)^{-1} \Phi^T D Y \\ &= \left(\Phi^T W \Phi\right)^{-1} \Phi^T W \left(\Phi w_{\mathrm{old}} + W^{-1} D Y\right) = \left(\Phi^T W \Phi\right)^{-1} \Phi^T W Z, \end{aligned}$$

with $Z = \Phi w_{\mathrm{old}} + W^{-1} D Y$.

Weighted least squares problem:

$$\sum_{i=1}^{n} \gamma_i (Y_i - \langle w, \Phi(X_i)\rangle)^2 = \langle Y - \Phi w, W(Y - \Phi w)\rangle ,$$

where $W = \mathrm{diag}(\gamma)$ and solution $w^* = \left(\Phi^T W \Phi\right)^{-1} \Phi^T W Y$.
Each update is the solution of a weighted least squares with output Z

iteratively reweighted least squares

Logistic Regression V

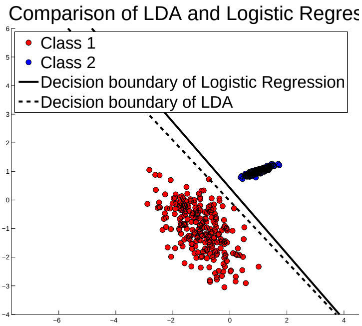
Problem: Empirical risk minimization is prone to overfitting.
Solution: Add regularizer !

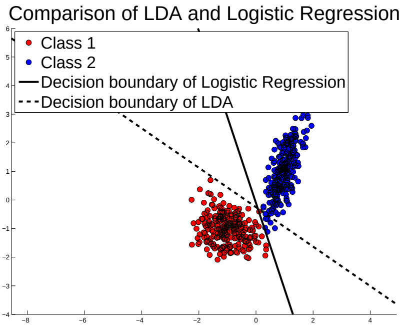
Definition

Given a training sample $T_n = (X_i, Y_i)_{i=1}^n$ with $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$ and the function space $\mathcal{F} = \{\sum_{i=1}^D w_i \phi_i(x) \,|\, w \in \mathbb{R}^D\}$ we define **L_2-regularized logistic regression** as the mapping $\mathcal{A} : T_n \to \mathcal{F}$ with,

$$T_n \mapsto f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \log\Big(1 + \exp(-Y_i \, \langle w, \phi(X_i)\rangle)\Big) + \lambda \, \|w\|_2^2 \,,$$

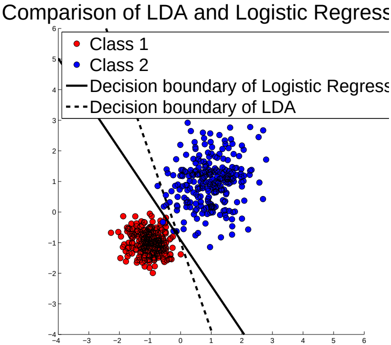
where λ is the regularization parameter.

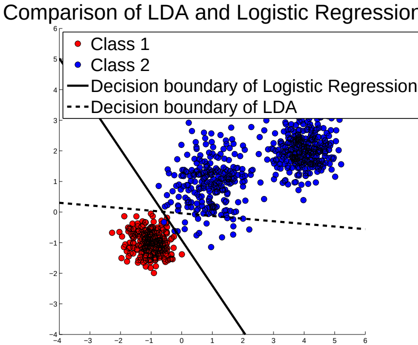




Left: A linearly separable problem, **Right:** A non-separable problem.

Comparison LDA vs Logistic Regression II





Left: Original data, **Right:** Adding the second Gaussian blob should not change the decision boundary. However, LDA changes its decision completely.

Support Vector Machines

The linear **support vector machine** can be motivated from different perspectives.

Geometric Perspective: Maximum margin hyperplane

Unique hyperplane which correctly classifies the data and has maximal distance/margin from the training data.

- **hard margin** case: linearly separable data.
- **soft margin** case: all kind of data allowed.

- Linear classifier is determined by the weight vector w and the offset b.

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

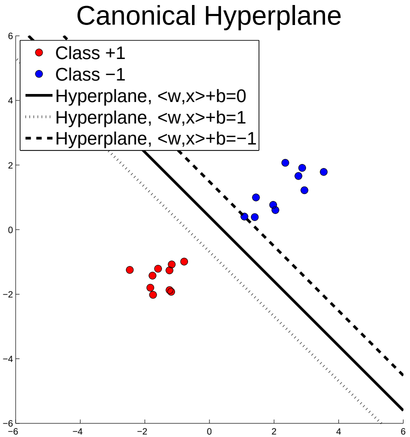
- **decision boundary** $\langle w, x\rangle + b = 0$ is the most interesting quantity.
- classifier and the decision boundary are not unique. For $\gamma > 0$, $\tilde{w} = \gamma w$ and $\tilde{b} = \gamma b$ gives same classifier. $\Rightarrow$ fix using **canonical hyperplane**.

Definition

The pair $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ is said to be in **canonical** form with respect to $X_1, \ldots, X_n \in \mathbb{R}^d$, if it scaled such that

$$\min_{i=1,\ldots,n} |\langle w, X_i\rangle + b| = 1,$$

which implies that the point closest to the hyperplane $h = \{x | \langle w, x\rangle + b = 0\}$ has distance $\rho = \frac{1}{\|w\|}$. We call ρ the **geometrical margin** of the hyperplane.



The canonical hyperplane for a set of training points $(X_i)_{i=1}^n$.

Support Vector Machines IV

Maximum margin hyperplane: a hyperplane which correctly classifies the data and has maximum distance/margin to the data.

Definition

A **maximum margin hyperplane** (w, b) for a **linearly separable** set of training data $(X_i, Y_i)_{i=1}^n$ is defined as

$$\max_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \min\{\|x - X_i\| \mid \langle w, x\rangle + b = 0,\, x \in \mathbb{R}^d,\; i = 1, \ldots, n\},$$

where we optimize over all (w, b) such that $Y_i\,(\langle w, X_i\rangle + b) > 0$.

Equivalent formulation:

$$\max_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \frac{1}{\|w\|}$$
$$\text{subject to: } Y_i(\langle w, X_i\rangle + b) \geq 1, \quad \forall\, i = 1, \ldots, n$$

Second equivalent formulation:

$$\min_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \frac{1}{2}\|w\|^2$$
$$\text{subject to: } Y_i(\langle w, X_i\rangle + b) \geq 1, \quad \forall\, i = 1, \ldots, n$$

- convex optimization problem: quadratic program

Lagrange function: Let $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^n$

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \Big[1 - Y_i(\langle w, X_i \rangle + b) \Big],$$

where $\alpha_i \geq 0, \quad \forall\, i = 1, \ldots, n$, are the **Lagrange multipliers**.

Dual Lagrange function:

$$q(\alpha) = \inf_{w \in \mathbb{R}^d,\ b \in \mathbb{R}} L(w, b, \alpha).$$

- since L is convex we can compute the dual using the stationary point,
- Slater condition fulfilled if data is linearly separable $\Rightarrow$ strong duality, we can solve primal problem via the dual problem.

Support Vector Machines VII

Derivatives:

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i Y_i X_i, \qquad \frac{\partial L(w,b,\alpha)}{\partial b} = -\sum_{i=1}^{n} \alpha_i Y_i.$$

Conditions for global minimum:

$$w = \sum_{i=1}^{n} \alpha_i Y_i X_i, \qquad \sum_{i=1}^{n} \alpha_i Y_i = 0.$$

Plugging these expressions into $L(w,b,\alpha)$ we get the dual Lagrangian

$$q(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle + \sum_{i=1}^{n} \alpha_i,$$

where $\alpha_i \geq 0, \quad \forall\, i = 1, \ldots, n$.

Dual problem:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle ,$$

$$\text{subject to: } \alpha_i \geq 0, \quad i = 1, \ldots, n,$$

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

- The dual problem is solved in practice using the SMO (Sequential minimal optimization) method.
- complexity is in the worst case cubic in n but often much faster.

Karush-Kuhn-Tucker (KKT) conditions: The most important one is the complementary slackness condition:

$$\begin{aligned} & \alpha_i > 0 \quad \text{if} \quad \Big[1 - Y_i(\langle w, X_i\rangle + b)\Big] = 0 \\ \text{and} \qquad & \alpha_i = 0 \quad \text{if} \quad \Big[1 - Y_i(\langle w, X_i\rangle + b)\Big] < 0. \end{aligned}$$

or more compactly

$$\alpha_i \Big[1 - Y_i(\langle w, X_i\rangle + b)\Big] = 0.$$

The offset b can thus be determined by averaging the value $b = Y_i - \langle w, X_i\rangle$ over all points with $\alpha_i > 0$:

$$b = \frac{1}{\sum_{i=1}^n \mathbb{1}_{\alpha_i > 0}} \sum_{i=1}^n \mathbb{1}_{\alpha_i > 0}\big(Y_i - \sum_{j=1}^n \alpha_j \, Y_j \, \langle X_i, X_j\rangle\big).$$

Final weight vector:

$$w = \sum_{i=1}^{n} \alpha_i Y_i X_i.$$

Only the points closest to the decision boundary contribute to solution

$$\alpha_i > 0 \quad \Leftrightarrow \quad \Big[1 - Y_i(\langle w, X_i\rangle + b)\Big] = 0,$$

These points are called **support vectors**. The area between the two supporting hyperplanes $\{x \mid \langle w, x\rangle + b = 1\}$ and $\{x \mid \langle w, x\rangle + b = -1\}$ is called the **margin**.

1. The weight vector of the support vector machine is typically **sparse** in terms of α.
2. Modifications of the training points matter only if they move into the margin.

Convex hull formulation

Equivalent reformulation of the dual problem:

$$\min_{\alpha \in \mathbb{R}^n} \Big\| \sum_{i=1, Y_i=1}^{n} \alpha_i X_i - \sum_{j=1, Y_j=-1}^{n} \alpha_j X_j \Big\|^2,$$

$$\text{subject to: } \alpha_i \geq 0, \quad i = 1, \ldots, n,$$

$$\sum_{i=1, Y_i=1}^{n} \alpha_i = \sum_{j=1, Y_j=-1}^{n} \alpha_j = 1.$$

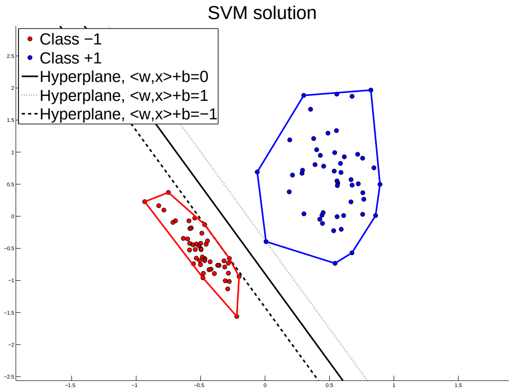
$\Longrightarrow$ distance between the convex hulls of positive and negative class.

Definition

Given a set T of points $(X_i)_{i=1}^n$ in $\mathbb{R}^d$. The **convex hull** of T is defined as the set

$$\Big\{ \sum_{i=1}^{n} \alpha_i X_i \mid \sum_{i=1}^{n} \alpha_i = 1, \quad \alpha_i \geq 0, \quad i = 1, \ldots, n \Big\}.$$

Example: linearly separable case



A linearly separable problem. The hard margin solution of the SVM is shown together with the convex hulls of the positive and negative class. The points on the margin, that is $\langle w, x \rangle + b = \pm 1$, are called **support vectors**.

Problems of the hard margin case:

- not all data is linearly separable,
- the **hard margin** case is often too strict since it is sensitive to outliers.

Relaxation of the constraints:

$$Y_i(\langle w, X_i \rangle + b) \geq 1 - \xi_i$$

where $\xi_i \geq 0$ are the **slack variables**.

Primal problem of the soft-margin case:

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

$$\text{subject to: } Y_i(\langle w, X_i \rangle + b) \geq 1 - \xi_i, \quad \forall\, i = 1, \ldots, n,$$

$$\xi_i \geq 0, \quad \forall\, i = 1, \ldots, n$$

Soft Margin as RERM

At the optimum: with $\xi_i \geq 0$,

$$\xi_i = \max\Big(0, 1 - Y_i(\langle w, X_i\rangle + b)\Big).$$

With $f(X_i) = \langle w, X_i\rangle + b$ we note that $\max\Big(0, 1 - y_i\, f(X_i)\Big)$ is nothing else than the **hinge loss**.

Soft Margin SVM is RERM with Hinge loss and L_2-regularization:

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

Error parameter C is inverse to the regularization parameter $\lambda = \frac{1}{C}$

Lagrangian of Soft Margin

Lagrangian of the soft margin problem:

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\left\|w\right\|^2 + \frac{C}{n}\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i \Big[1 - \xi_i - Y_i(\langle w, X_i\rangle + b)\Big] - \sum_{i=1}^n \beta_i \xi_i$$

where $\alpha_i \geq 0,\ i = 1,\ldots,n$ and $\beta_i \geq 0,\ i = 1,\ldots,n$.

Conditions for stationary point:

$$w = \sum_{i=1}^n \alpha_i Y_i X_i, \qquad \sum_{i=1}^n \alpha_i Y_i = 0, \qquad \beta = \frac{C}{n}\mathbb{1} - \alpha.$$

The last equation can be used to get rid of β. Due to the positivity of β we get the new constraint for α

$$0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1,\ldots,n.$$

Dual Lagrangian of the soft margin problem:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle \,,$$

$$\text{subject to: } 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^{n} Y_i \alpha_i = 0.$$

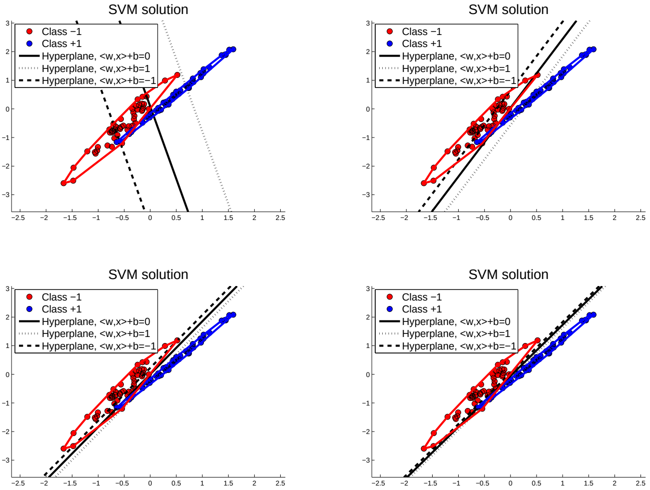
Complementary slackness conditions (part of KKT conditions):

$$\alpha_i \Big[1 - \xi_i - Y_i(\langle w, X_i \rangle + b) \Big] = 0, \qquad \text{and} \qquad \beta_j \xi_j = 0, \qquad i, j = 1, \ldots, n.$$

Three classes of points:

- $\alpha_i = 0$: outside the margin and all correctly classified.
- $0 < \alpha_i < \frac{C}{n}$: lie exactly on the margin are all correctly classified.
- $\alpha_i = \frac{C}{n}$: inside the margin, can be misclassified.

Comparison of different C



Top row: error parameter $C = 10^1$ (left) and $C = 10^2$ (right), Bottom row: error parameter $C = 10^3$ (left) and $C = 10^4$ (right).