The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Convex Optimization — Fall 2019

LECTURE 23

Aryan Mokhtari Wednesday, November 20, 2019.

Goal: In this lecture, we first talk about the classification problem and in particular, we talk about linear discrimination and support vector machines.

1 Classification

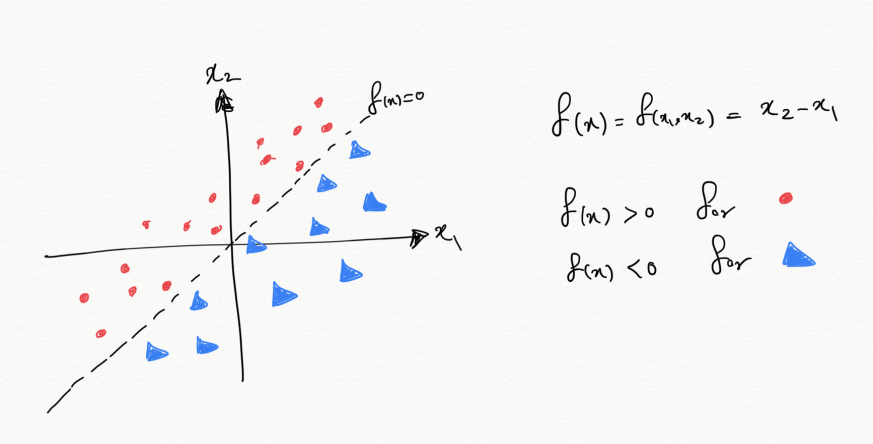
In classification problems,

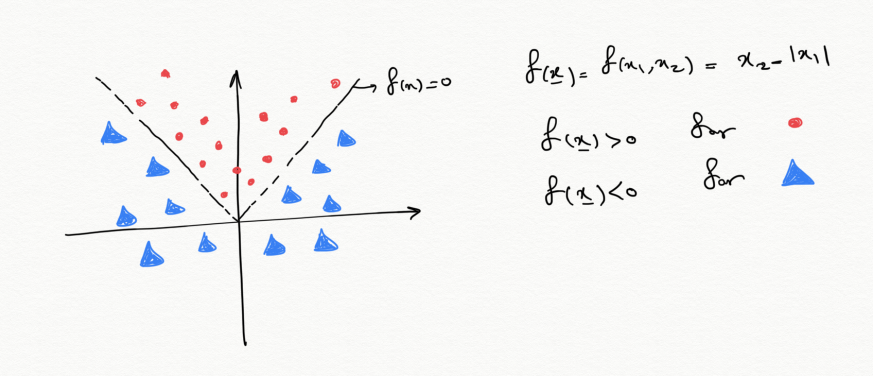
1. we are given two set of data points in $\mathbb{R}^n$ with different labels
2. we aim to find a function $f : \mathbb{R}^n \to \mathbb{R}$ that separates the data points, i.e., it is positive for the first group and negative for the second group.

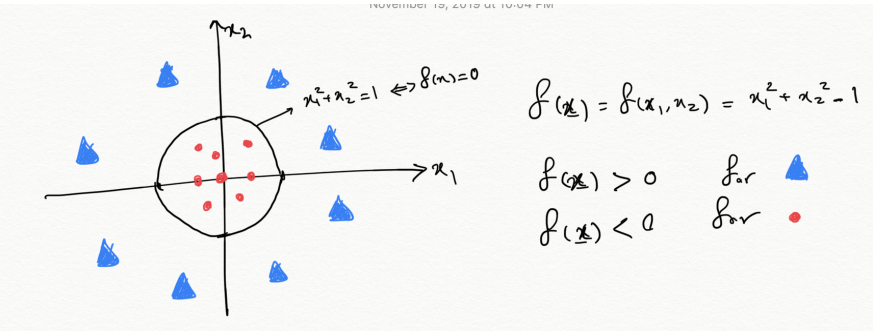
Consider the case that we have m samples and k of them have label 1 (circles) and $m-k$ have label -1 (triangles). The goal is to find a function that separates the samples with different labels. WLOG, assume that the labels y_i for $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are 1 and the labels for $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_m$ are -1. Then, the goal is to find f such that

$$f(\mathbf{x}_i) > 0, \quad \text{for } i = 1, \ldots, k \qquad\qquad f(\mathbf{x}_i) < 0, \quad \text{for } i = k+1, \ldots, m$$

If we can find such function, then we say that f *or its* 0*-level set* $\{\mathbf{x} | f(\mathbf{x}) = 0\}$ *separates (classifies) the two sets of points.*







1.1 Linear discrimination

In the linear discrimination we look for an affine function $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} - b$ that separates the datasets, i.e.,

$$\mathbf{a}^\top \mathbf{x}_i - b > 0, \quad \text{for } i = 1, \ldots, k \qquad\qquad \mathbf{a}^\top \mathbf{x}_i - b < 0, \quad \text{for } i = k+1, \ldots, m$$

The geometric interpretation of this classifies is that we are looking for **hyperplane** that separates the given datasets.
Now we can write this problem as a feasibility problem:

$$\begin{array}{ll} \text{maximize} & 0 \\ \text{subject to} & \mathbf{a}^\top \mathbf{x}_i - b > 0, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\ & \mathbf{a}^\top \mathbf{x}_i - b < 0, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \end{array}$$

This is a convex feasibility problem (in fact a set of linear inequalities). It may or may not be feasible.

Q: Can you identify the geometric condition for feasibility of linear classification?
A: Convex hull of samples with different labels do not intersect! (can be proven using the theorem of alternatives.)

If there exists an affine function that separates the data points with different labels, then *we can choose one that optimizes some measure of robustness.*

Goal: We are looking for a hyperplane that separates data points with different labels and has the maximum margin (maximum distance to any points in both datasets).

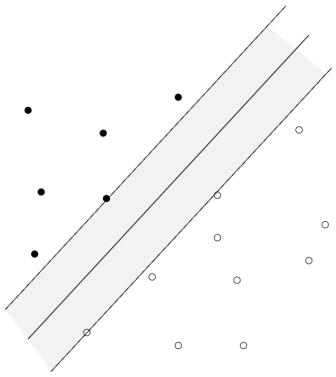
Geometric interpretation: Among all possible slabs (illustrated in the picture), we want to choose the one that has the largest width. In this case the margin is defined as the width of the slab, and the separating hyperplane is the one that is exactly in the middle of two boundaries of the slab.

If the separating hyperplane is $\mathbf{a}^\top \mathbf{x} - b = 0$, then we can always consider the ones on the boundary of the slab to be $\mathbf{a}^\top \mathbf{x} - b = 1$ and $\mathbf{a}^\top \mathbf{x} - b = -1$. (Here instead of 1 and -1 we can use any other positive value ϵ and its negative $-\epsilon$ by rescaling $\mathbf{a}$ and b)

(Euclidean) distance between hyperplanes

$$\begin{aligned} \mathcal{H}_1 &= \{z \mid a^T z + b = 1\} \\ \mathcal{H}_2 &= \{z \mid a^T z + b = -1\} \end{aligned}$$

is $\mathbf{dist}(\mathcal{H}_1, \mathcal{H}_2) = 2/\|a\|_2$



Hence, the problem that we want to solve is

$$\begin{aligned} \max_{\mathbf{a},b} \quad & \frac{2}{\|\mathbf{a}\|_2} \\ \text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq 1, \quad \text{for } y_i = 1 \\ & \mathbf{a}^\top \mathbf{x}_i - b \leq -1, \quad \text{for } y_i = -1 \end{aligned}$$

where the objective function is the width of the slab and the constraints ensure that the samples are correctly classified.

Side note: The distance between a point $\mathbf{x}_0$ and a hyperplane $\mathbf{a}^\top \mathbf{x} - b = 0$ is given by $\frac{|\mathbf{a}^\top \mathbf{x}_0 - b|}{\|\mathbf{a}\|_2}$. Moreover, the distance between two parallel hyperplanes $\mathbf{a}^\top \mathbf{x} - b = 0$ and $\mathbf{a}^\top \mathbf{x} - \hat{b} = 0$ is $\frac{|b - \hat{b}|}{\|\mathbf{a}\|_2}$.

Direct derivation of Max-margin problem: If we plan to maximize the distance between the separating hyperplane and all the samples points, then we should solve the following optimization problem:

$$\begin{aligned} \max_{\mathbf{a},b} \quad & \min_{i=1,\ldots,m} \left\{ \frac{|\mathbf{a}^\top \mathbf{x}_1 - b|}{\|\mathbf{a}\|_2}, \ldots, \frac{|\mathbf{a}^\top \mathbf{x}_m - b|}{\|\mathbf{a}\|_2} \right\} \\ \text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b > 0, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\ & \mathbf{a}^\top \mathbf{x}_i - b < 0, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \end{aligned}$$

Margin is the minimum distance to any of the points in the dataset times two.

We can also rewrite this problem as

$$\begin{aligned}
\max_{\mathbf{a},b,t} \quad & \frac{t}{\|\mathbf{a}\|_2} \\
\text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b > 0, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \mathbf{a}^\top \mathbf{x}_i - b < 0, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\
& |\mathbf{a}^\top \mathbf{x}_i - b| \geq t, \quad \text{for } i = 1, \ldots, m \\
& t > 0.
\end{aligned}$$

which can be written as

$$\begin{aligned}
\max_{\mathbf{a},b,t} \quad & \frac{t}{\|\mathbf{a}\|_2} \\
\text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b > 0, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \mathbf{a}^\top \mathbf{x}_i - b < 0, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\
& \mathbf{a}^\top \mathbf{x}_i - b \geq t, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \mathbf{a}^\top \mathbf{x}_i - b \leq -t, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\
& t > 0.
\end{aligned}$$

which can be simplified as

$$\begin{aligned}
\max_{\mathbf{a},b,t} \quad & \frac{t}{\|\mathbf{a}\|_2} \\
\text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq t, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \mathbf{a}^\top \mathbf{x}_i - b \leq -t, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\
& t > 0.
\end{aligned}$$

We can eliminate the variable t by the following change of variables $\hat{\mathbf{a}} = \mathbf{a}/t$ and $\hat{b} = b/t$

$$\begin{aligned}
\max_{\hat{\mathbf{a}},\hat{b}} \quad & \frac{1}{\|\hat{\mathbf{a}}\|_2} \\
\text{s. t.} \quad & \hat{\mathbf{a}}^\top \mathbf{x}_i - \hat{b} \geq 1, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \hat{\mathbf{a}}^\top \mathbf{x}_i - \hat{b} \leq -1, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1)
\end{aligned}$$

The above problem is called the max-margin linear classification problem. To write it as a convex minimization we rewrite as

$$\begin{aligned}
\min_{\mathbf{a},b} \quad & \|\mathbf{a}\|_2 \\
\text{s. t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq 1, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\
& \mathbf{a}^\top \mathbf{x}_i - b \leq -1, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1)
\end{aligned}$$

(after squaring objective) a QP in $\mathbf{a}, b$.
Samples on the margin are called the support vectors. We can also obtain the support vector by

solving the dual problem. The dual function can be written as

$$
\begin{aligned}
g(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) &= \min_{\mathbf{a},b} \left(\|\mathbf{a}\|_2 + \sum_{i=1}^{k} \lambda_i (1 - \mathbf{a}^\top \mathbf{x}_i + b) + \sum_{i=k+1}^{m} \hat{\lambda}_i (\mathbf{a}^\top \mathbf{x}_i - b + 1) \right) \\
&= \sum_{i=1}^{k} \lambda_i + \sum_{i=k+1}^{m} \hat{\lambda}_i + \min_{\mathbf{a},b} \left(\|\mathbf{a}\|_2 - \mathbf{a}^\top \Big(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i - \sum_{i=k+1}^{m} \hat{\lambda}_i \mathbf{x}_i \Big) + b \Big(\sum_{i=1}^{k} \lambda_i - \sum_{i=k+1}^{m} \hat{\lambda}_i \Big) \right) \\
&= \begin{cases} \sum_{i=1}^{k} \lambda_i + \sum_{i=k+1}^{m} \hat{\lambda}_i & \left\| \sum_{i=1}^{k} \lambda_i \mathbf{x}_i - \sum_{i=k+1}^{m} \hat{\lambda}_i \mathbf{x}_i \right\| \leq 1, \quad \sum_{i=1}^{k} \lambda_i - \sum_{i=k+1}^{m} \hat{\lambda}_i = 0 \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}
$$

where $\boldsymbol{\lambda} \in \mathbb{R}^k$ and $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{m-k}$. Therefore, the dual problem is

$$
\begin{aligned}
\max_{\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}} \quad & \mathbf{1}^\top \boldsymbol{\lambda} + \mathbf{1}^\top \hat{\boldsymbol{\lambda}} \\
\text{s. t.} \quad & \left\| \sum_{i=1}^{k} \lambda_i \mathbf{x}_i - \sum_{i=k+1}^{m} \hat{\lambda}_i \mathbf{x}_i \right\| \leq 1 \\
& \mathbf{1}^\top \boldsymbol{\lambda} = \mathbf{1}^\top \hat{\boldsymbol{\lambda}}, \qquad \boldsymbol{\lambda} \geq \mathbf{0}, \ \hat{\boldsymbol{\lambda}} \geq \mathbf{0}
\end{aligned}
$$

Using the following change of variables:

$$
\theta_i = \frac{\lambda_i}{\mathbf{1}^\top \boldsymbol{\lambda}}, \quad \gamma_i = \frac{\hat{\lambda}_i}{\mathbf{1}^\top \boldsymbol{\lambda}}, \quad t = \frac{1}{\mathbf{1}^\top \boldsymbol{\lambda} + \mathbf{1}^\top \hat{\boldsymbol{\lambda}}}
$$

we can rewrite the problem as

$$
\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\gamma}, t} \quad & t \\
\text{s. t.} \quad & \frac{1}{2} \left\| \sum_{i=1}^{k} \theta_i \mathbf{x}_i - \sum_{i=k+1}^{m} \gamma_i \mathbf{x}_i \right\|_2 \leq t \\
& \mathbf{1}^\top \boldsymbol{\theta} = 1, \quad \mathbf{1}^\top \boldsymbol{\gamma} = 1, \qquad \boldsymbol{\theta} \geq \mathbf{0}, \ \boldsymbol{\gamma} \geq \mathbf{0}
\end{aligned}
$$

which is

$$
\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\gamma}} \quad & \frac{1}{2} \left\| \sum_{i=1}^{k} \theta_i \mathbf{x}_i - \sum_{i=k+1}^{m} \gamma_i \mathbf{x}_i \right\|_2 \\
\text{s. t.} \quad & \mathbf{1}^\top \boldsymbol{\theta} = 1, \quad \mathbf{1}^\top \boldsymbol{\gamma} = 1, \qquad \boldsymbol{\theta} \geq \mathbf{0}, \ \boldsymbol{\gamma} \geq \mathbf{0}
\end{aligned}
$$

The objective function finds the (1/2) of the minimum distance between the convex hull of the samples with label 1 and the convex hull of the samples with label -1. The nonzero multipliers θ_i^* and γ_i^* indicate the points at the margin for each of the data sets.

1.2 Data that is not linearly separable

In this case, one might sat that we should try to find the classifier that minimizes the number of misclassified samples.

This problem can be written as

$$\begin{aligned} \min_{\mathbf{a},b,\boldsymbol{\eta},\boldsymbol{\zeta}} \quad & \|\boldsymbol{\eta}\|_0 + \|\boldsymbol{\zeta}\|_0 \\ \text{s.t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq 1 - \eta_i, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\ & \mathbf{a}^\top \mathbf{x}_i - b \leq -1 + \zeta_i, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\ & \boldsymbol{\eta} \geq \mathbf{0}, \quad \boldsymbol{\zeta} \geq \mathbf{0} \end{aligned}$$

Here, $\|\boldsymbol{\eta}\|_0$ denotes the number of nonzero elements of $\boldsymbol{\eta}$. Unfortunately, it is a hard problem to solve and we need to relax it:

$$\begin{aligned} \min_{\mathbf{a},b,\boldsymbol{\eta},\boldsymbol{\zeta}} \quad & \|\boldsymbol{\eta}\|_1 + \|\boldsymbol{\zeta}\|_1 \\ \text{s.t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq 1 - \eta_i, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\ & \mathbf{a}^\top \mathbf{x}_i - b \leq -1 + \zeta_i, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\ & \boldsymbol{\eta} \geq \mathbf{0}, \quad \boldsymbol{\zeta} \geq \mathbf{0} \end{aligned}$$

More generally, we can consider the trade-off between the number of misclassified points, and the width of the slab which is proportional to $1/\|\mathbf{a}\|_2$. The standard support vector classifier for the sets $\mathbf{x}_1, \ldots, \mathbf{x}_m$ is defined as

$$\begin{aligned} \min_{\mathbf{a},b,\boldsymbol{\eta},\boldsymbol{\zeta}} \quad & \|\mathbf{a}\|_2^2 + \gamma(\|\boldsymbol{\eta}\|_1 + \|\boldsymbol{\zeta}\|_1) \\ \text{s.t.} \quad & \mathbf{a}^\top \mathbf{x}_i - b \geq 1 - \eta_i, \quad \text{for } i = 1, \ldots, k \text{ (for } y_i = 1) \\ & \mathbf{a}^\top \mathbf{x}_i - b \leq -1 + \zeta_i, \quad \text{for } i = k+1, \ldots, m \text{ (for } y_i = -1) \\ & \boldsymbol{\eta} \geq \mathbf{0}, \quad \boldsymbol{\zeta} \geq \mathbf{0} \end{aligned}$$

We can also write down this problem in a compressed form as

$$\begin{aligned} \min_{\mathbf{a},b,\boldsymbol{\eta}} \quad & \frac{1}{m}\sum_{i=1}^{m} \eta_i + \lambda\|\mathbf{a}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{a}^\top \mathbf{x}_i - b) \geq 1 - \eta_i, \quad \text{for } i = 1, \ldots, m \\ & \eta_i \geq 0 \quad \text{for } i = 1, \ldots, m \end{aligned}$$

which is

$$\begin{aligned} \min_{\mathbf{a},b,\boldsymbol{\eta}} \quad & \frac{1}{m}\sum_{i=1}^{m} \eta_i + \lambda\|\mathbf{a}\|_2^2 \\ \text{s.t.} \quad & 1 - y_i(\mathbf{a}^\top \mathbf{x}_i - b) \leq \eta_i, \quad \text{for } i = 1, \ldots, m \\ & 0 \leq \eta_i \quad \text{for } i = 1, \ldots, m \end{aligned}$$

and can be simplified as the following unconstrained problem

$$\min_{\mathbf{a},b,\boldsymbol{\eta}} \quad \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i(\mathbf{a}^\top \mathbf{x}_i - b)\} + \lambda\|\mathbf{a}\|_2^2$$

which is known as Soft-margin SVM.