Operations that preserve convexity

Source: Ahmadi, ORF 363 slides.

Review

- Convex optimization problems have several nice properties
- Every local minimum is a global minimum
- If the objective function is convex, there is at most one global minimum

Review

How to check that a function f is convex?

- $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y, \forall \lambda \in [0,1]$

Review

How to check that a function f is convex?

- $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y, \forall \lambda \in [0,1]$
- $f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall x, y \in \mathsf{dom}(f)$
- $\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathsf{dom}(f)$

Overview

- In general, testing convexity can be intractable
- it is useful to produce as many convex functions as we can from a ground set of functions that we already know are convex.

Overview

- In general, testing convexity can be intractable
- it is useful to produce as many convex functions as we can from a ground set of functions that we already know are convex.
- We will talk about some operations that take convex functions as input and produce more convex function.

Nonnegative weighted sums

Composition with an affine mapping

Pointwise maximum

Restriction to a line

Nonnegative weighted sums

If $f_1, f_2, \ldots, f_n$ are convex functions and $\omega_1, \omega_2, \ldots, \omega_n \geq 0$, then

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x) + \ldots + \omega_n f_n(x)$$

is also convex

Nonnegative weighted sums

If $f_1, f_2, \ldots, f_n$ are convex functions and $\omega_1, \omega_2, \ldots, \omega_n \geq 0$, then

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x) + \ldots + \omega_n f_n(x)$$

is also convex

Similarly, a nonnegative weighted sum of concave functions in also concave

Nonnegative weighted sums

Let $f_1, f_2, \ldots, f_n$ are convex functions and $\omega_1, \omega_2, \ldots, \omega_n \geq 0$ and

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x) + \ldots + \omega_n f_n(x)$$

Remarks

- In particular, $\alpha \geq 0$ and f is convex $\Rightarrow \alpha f$ is convex
- If f_1 and f_2 are convex $\Rightarrow f_1 + f_2$ is convex

Remarks

- In particular, $\alpha \geq 0$ and f is convex $\Rightarrow \alpha f$ is convex
- If f_1 and f_2 are convex $\Rightarrow f_1 + f_2$ is convex
- From this, we can have more complex constraints e.g.
  - Convex function $\leq$ Concave function

Some questions

If f_1 and f_2 are convex functions,

- is $f_1 - f_2$ convex?
- is $f_1 \times f_2$ convex?
- is $\frac{f_1}{f_2}$ convex?

Nonnegative weighted sums

Composition with an affine mapping

Pointwise maximum

Restriction to a line

Composition with an affine mapping

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$, then the function

$$g(x) = f(Ax + b)$$

is also convex

Composition with an affine mapping

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$, then the function

$$g(x) = f(Ax + b)$$

is also convex

If f is concave then so is g

Composition with an affine mapping

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^m$, then the function

$$g(x) = f(Ax + b)$$

Example

$$f(x_1, x_2) = (x_1 - 2x_2)^4 + 2e^{3x_1+2x_2-5}$$

Nonnegative weighted sums

Composition with an affine mapping

Pointwise maximum

Restriction to a line

Pointwise maximum

If $f_1, f_2, \ldots, f_m$ are convex functions then their pointwise maximum

$$f(x) = \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$$

is also convex

Pointwise maximum

If $f_1, f_2, \ldots, f_m$ are convex functions then their pointwise maximum

$$f(x) = \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$$

Example

The hinge loss: $g(x) = \max\{0, 1 - x\}$

Remarks

Pointwise minimum of two concave functions is also concave

But the pointwise minimum of two convex functions may not be convex

Nonnegative weighted sums

Composition with an affine mapping

Pointwise maximum

Restriction to a line

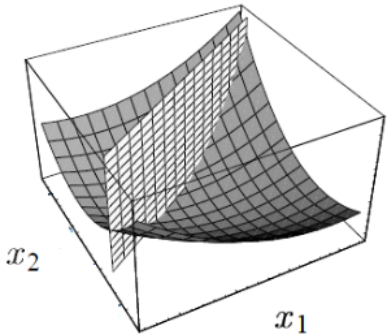
Restriction to a line

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and fix some $x, y \in \mathbb{R}^n$. Then, the function $g : \mathbb{R} \to \mathbb{R}$ given by

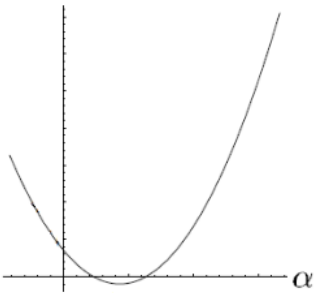
$$g(t) = f(x + ty)$$

is convex

Restriction to a line



$f(x_1, x_2)$



$f(\alpha)$

Solving convex optimization problems in CVXPY

Problem 1

Variables: $x_1, x_2 \in \mathbb{R}$

$$\min_{x_1, x_2} \quad 140x_1 + 235x_2$$

subject to

$$\begin{aligned} x_1 &\geq 0 \\ x_2 &\geq 0 \\ x_1 + x_2 &\leq 180 \\ x_1 + 2x_2 &\leq 240 \\ 0.3x_1 + 0.1x_2 &\leq 30 \\ x_1^2 + 2x_2^2 &\leq 1 \end{aligned}$$

Problem 2

$A \in \mathbb{R}^{16\times 8}, b \in \mathbb{R}^{16}$

Variables: $x \in \mathbb{R}^8$

$$\min_x \ \|Ax - b\|^2$$

Problem 3

$$A \in \mathbb{R}^{16\times 8}, b \in \mathbb{R}^{16}$$

Variables: $x \in \mathbb{R}^8$

$$\min_x \ \|Ax - b\|^2$$

subject to

$$x_1, \ldots, x_8 \geq -20$$

$$\|x\|_\infty \leq 100$$

Problem 4

Variables: $x_1, x_2 \in \mathbb{R}$

$$\min_{x_1, x_2} \quad x_1 + x_2$$

subject to

$$x_1 + x_2 \leq -5$$
$$x_1^2 + x_2^2 \leq 1$$

Applications of Convex Optimization

LASSO

Support Vector Machines (SVMs)

Personalized pricing of flight tickets

- You have been hired as a quantitative analyst by Priceline.com, a major travel website company

- you want an efficient mechanism for predicting the highest price a given customer is willing to pay for a flight.

Data

As usual in the age of big data and no privacy, your boss gives you a massive data set containing information about past customers:

- Age
- number of friends they have on Facebook
- The place where they currently live
- How frequently they travel
- Their average monthly salary
- Gender
- Marital status
- Average time spent on the internet
- How many times they searched for the same flight
- **price they paid for the ticket**

Problem formulation

Overall, you have $p+1$ vectors in $\mathbb{R}^n$:

$$x_1 = \begin{pmatrix} x_1^1 \\ x_1^2 \\ \vdots \\ x_1^n \end{pmatrix}, x_2 = \begin{pmatrix} x_2^1 \\ x_2^2 \\ \vdots \\ x_2^n \end{pmatrix}, \ldots, x_p = \begin{pmatrix} x_p^1 \\ x_p^2 \\ \vdots \\ x_p^n \end{pmatrix}, \quad y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{pmatrix}$$

Problem formulation

You would like to find a simple relationship between and the attiributes. For each customers you believe that:

$$y^j \approx \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{monthly salary} + \dots$$

Problem formulation

You would like to find a simple relationship between and the attiributes. For each customers you believe that:

$$y^j \approx \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{monthly salary} + \dots$$

or in our notations,

$$y^j \approx \beta_0 + \beta_1 x_1^j + \beta_2 x_2^j + \dots + \beta_p x_p^j, \quad \forall j$$

Problem formulation

You would like to find a simple relationship between and the attiributes. For each customers you believe that:

$$y^j \approx \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{monthly salary} + \dots$$

or in our notations,

$$y^j \approx \beta_0 + \beta_1 x_1^j + \beta_2 x_2^j + \dots + \beta_p x_p^j, \quad \forall j$$

In vector notation,

$$y \approx \beta_0 \mathbf{1} + \sum_{k=1}^{p} \beta_k x_k \Rightarrow y - \beta_0 \mathbf{1} - \sum_{k=1}^{p} \beta_k x_k \approx 0$$

Problem formulation

The natural optimization problem to solve to find the best coefficients ($\beta_0, \ldots, \beta_p$) is then:

$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2$$

This is a convex optimization problem. Which rule would you use to see this?

Removing redundant attributes

- You also feel that many of the attributes are irrelevant, so you want to find a solution where many β_k's are zero; i.e., the behavior of the customors is explained by a few attributes only

Removing redundant attributes

- You also feel that many of the attributes are irrelevant, so you want to find a solution where many β_k's are zero; i.e., the behavior of the customors is explained by a few attributes only

- A common approach is to make

$$\|\beta\|_1 = |\beta|_1 + |\beta_2| + \ldots + |\beta_p|$$

  as small as possible. This encourages many β_k's to be zero or close to zero

LASSO problem

Your new optimization problem then becomes:

$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2 + \gamma \|\beta\|_1$$

where γ is some positive constant picked by you. This is again a convex optimization problem (why?)

LASSO problem

$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2 + \gamma \|\beta\|_1$$

- If γ is big, then more effort goes into minimizing the second term

LASSO problem

$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2 + \gamma \|\beta\|_1$$

- If γ is big, then more effort goes into minimizing the second term
- If γ is small, then more effort goes into minimizing the first term

LASSO problem

$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2 + \gamma \|\beta\|_1$$

- If γ is big, then more effort goes into minimizing the second term
- If γ is small, then more effort goes into minimizing the first term
- Making the first term small makes the error in the prediction of the model on past customers small

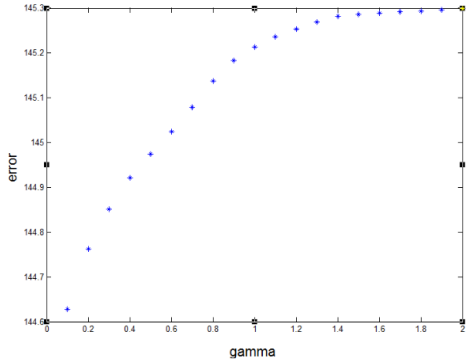
LASSO problem

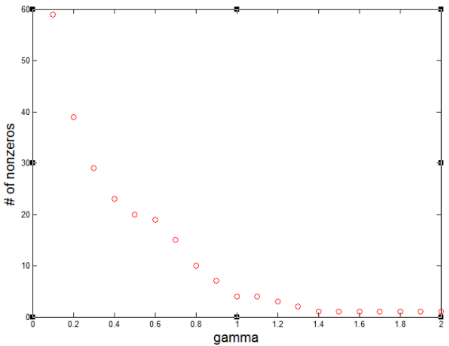
$$\min_{\beta_0,\ldots,\beta_p} \left\| y - \beta_0 - \sum_{i=1}^{p} \beta_k x_k \right\|_2 + \gamma \|\beta\|_1$$

- If γ is big, then more effort goes into minimizing the second term
- If γ is small, then more effort goes into minimizing the first term
- Making the first term small makes the error in the prediction of the model on past customers small
- Making the second term small ensures that we don't use too many attributes in our predictor (avoid overfitting)

Example

- We solve this problem with CVX with
  - $n = 10000$
  - $p = 100$
  - β between 0.1 and 2





LASSO

Support Vector Machines (SVMs)

Supervised learning

- Supervised learning: learn a classifier from a labeled data set (called the training set).
- The classifier is then used to label future data points
- Classic example is an email spam filter:
  - The emails that already have the labels constitute the ”training set”.

Spam or Not Spam?

Hello class,

My office hours this week have moved to Thursday, 4-5:30 PM. Lecture 4 is now up on the course website.

-Bom

Good day,

My name is Chaghal. I seek true soulmate. Are you ready for relations? Check my profile here:

http://soul4.com/me.exe

Hey man,

I'm tired of this homework for STAT 424. Let's go party tonight. We can always ask for an extension

-J

Feature vector

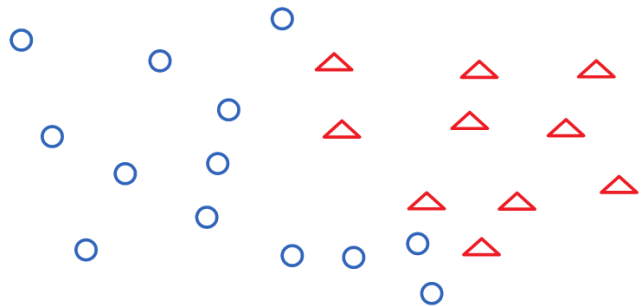
- A basic approach is to associate a pair (x_i, y_i) to each email;
  - y_i is the label: 1 (spam) or -1 (not spam)

Feature vector

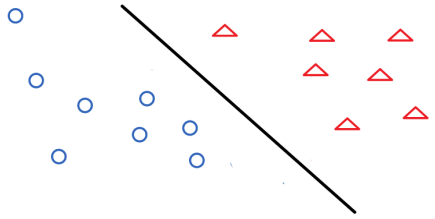
- A basic approach is to associate a pair (x_i, y_i) to each email;
  - y_i is the label: 1 (spam) or -1 (not spam)
  - $x_i \in \mathbb{R}^n$ is called a **feature vector**; it collects some relevant information about email i. For example,
    - How many words are in the email?
    - How many misspelled words?
    - How many links?
    - Is there a $ sign?
    - Does the word “bank account” appear?
    - Is the sender’s email client trustworthy?

Support vector machines

If we have m emails, we end up with m vectors in $\mathbb{R}^n$, each with a label $+1$ or -1. Here is a toy example in $\mathbb{R}^2$



Linear boundary



- The simplest one is linear classification: $f(x) = a^T x - b$

- Need to find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ that satisfy:

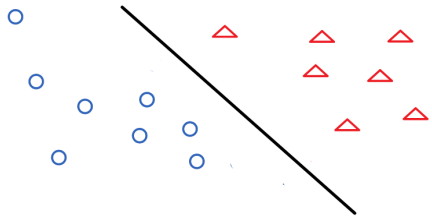
$$a^T x_i - b > 0 \quad \text{if } y_i = 1$$
$$a^T x_i - b < 0 \quad \text{if } y_i = -1$$

Linear boundary

$$a^T x_i - b > 0 \quad \text{if } y_i = 1$$
$$a^T x_i - b < 0 \quad \text{if } y_i = -1$$

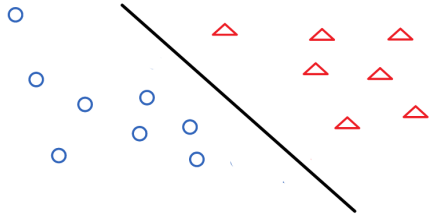
This is equivalent to:

$$y_i(a^T x_i - b) > 0, \quad i = 1, \ldots, m$$



We also want to **maximize** the margin: find a, b and some $t > 0$ such that

$$y_i(a^T x_i - b) \geq t, \quad i = 1, \ldots, m$$

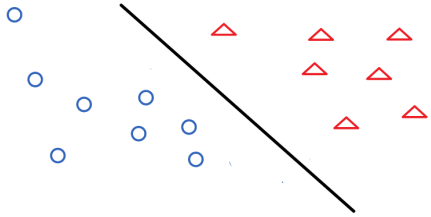


We also want to **maximize** the margin: find a, b such that

$$y_i(a^T x_i - b) \geq 1, \quad i = 1, \ldots, m$$

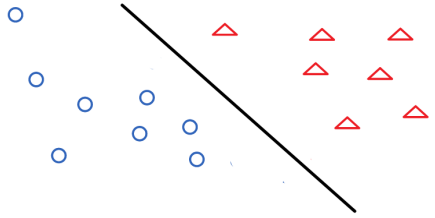
Fact: distance of a point $v \in \mathbb{R}^n$ to a hyperplane $a^T x - b = 0$ is

$$\frac{|a^T v - b|}{\|a\|}$$



$$\max_{a,b} \ \frac{1}{\|a\|}$$

$$\text{s.t. } y_i(a^T x_i - b) \geq 1, \quad i = 1, \ldots, m$$

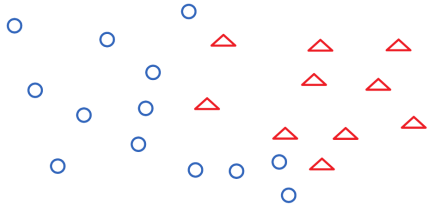


$$\max_{a,b} \ \frac{1}{\|a\|}$$
$$\text{s.t. } y_i(a^T x_i - b) \geq 1, \quad i = 1, \ldots, m$$

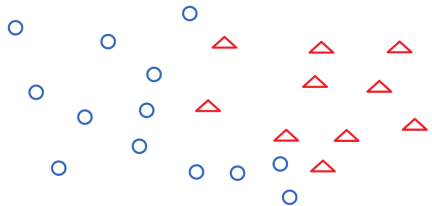
which is equivalent to

$$\min_{a,b} \ \|a\|$$
$$\text{s.t. } y_i(a^T x_i - b) \geq 1, \quad i = 1, \ldots, m$$

What is the points are not linearly separable?

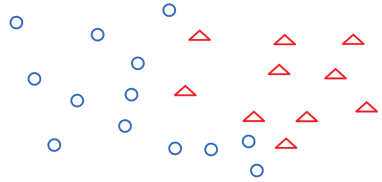


What is the points are not linearly separable?



Then we allow some points x_i to move at distance η_i across the margin:

$$\text{s.t. } y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m$$
$$\eta_i \geq 0, \quad i = 1, \ldots, m$$



But we don't want x_i to move too much i.e. we want to minimize η_i's

$$\begin{aligned} \min_{a,b} \quad & \|a\| \\ \text{s.t.} \quad & y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m \\ & \eta_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Soft-margin SVMs

A common approach to minimize η_i's is to solve the following problem:

$$\begin{aligned} \min_{a,b} \quad & \|a\| + \gamma \|\eta\|_1 \\ \text{s.t.} \quad & y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m \\ & \eta_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Soft-margin SVMs

$$\begin{aligned} \min_{a,b} \quad & \|a\| + \gamma \|\eta\|_1 \\ \text{s.t.} \quad & y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m \\ & \eta_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

- $\gamma \geq 0$ is a constant picked by you
- For each γ, this is a convex problem

Soft-margin SVMs

$$\begin{aligned} \min_{a,b} \quad & \|a\| + \gamma \|\eta\|_1 \\ \text{s.t.} \quad & y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m \\ & \eta_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

- $\gamma \geq 0$ is a constant picked by you
- For each γ, this is a convex problem
- Larger γ means we assign more importance to reducing number of misclassified points

Soft-margin SVMs

$$\begin{aligned} \min_{a,b} \quad & \|a\| + \gamma \|\eta\|_1 \\ \text{s.t.} \quad & y_i(a^T x_i - b) \geq 1 - \eta_i, \quad i = 1, \ldots, m \\ & \eta_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

- $\gamma \geq 0$ is a constant picked by you
- For each γ, this is a convex problem
- Larger γ means we assign more importance to reducing number of misclassified points
- Smaller γ means we assign more importance to having a large margin.