• introd. to nonlinear optimiz   AMIR BECK

CHAP 7. SECTION 7.4

TASK 7) Resolution:

show that $\tilde{x} = x - P\,\mathrm{sgn}(yw)$ solve $\underset{\tilde{x}}{\text{minimize}}\ y(w_0 + \tilde{x}^T w)$

sub. to $|\tilde{x}_d - x_d| \leq P$, for $1 \leq d \leq D$

each $\tilde{x}_d$ can only deviate from x_d by at most P in either direction.

sgn(·) acts comp. wise on vectors

(if $w = (u_1, u_2, u_3)$) => $\mathrm{sgn}(w) = (\mathrm{sgn}(u_1), \mathrm{sgn}(u_2), \mathrm{sgn}(u_3))$

$y \in \{-1, 1\}$ $\tilde{x}$: attacked feature vector

w: classifier parameter

P: maximum perturbation the attacker can introduce

Attacker goal is to minimize $y(w_0 + \tilde{x}^T w)$; we have to modify $\tilde{x}$ such that the classifier is more likely to make an incorrect decision

let's force $y(w_0 + \tilde{x}^T w)$ to be negative

$\mathrm{sgn}(y \cdot w) = (\mathrm{sgn}(y \cdot w_1), \mathrm{sgn}(y \cdot w_2), \ldots, \mathrm{sgn}(y \cdot w_D))$ since it acts a component-wise on vectors.

-> where each component indicates the direction (positive or negative) of the corresponding weigth w_d.

- change in $\tilde{x}^T w$: the inner product $\tilde{x}^T w = \sum_{d=1}^{D} \tilde{x}_d w_d$ depends on each component $\tilde{x}_d$. The optimal strategy for minimizing $y(w_0 + \tilde{x}^T w)$ is to adjust each component $\tilde{x}_d$ in a way that reduces the value of $\tilde{x}_d \cdot w_d$ as much as possible.
  - Maximizing the neg impact: The value $\tilde{x}_d \cdot w_d$ is minimized when $\tilde{x}_d$ takes the value farthest in the opposite direction of w_d;

    i.e. when $\tilde{x}_d = x_d - P \cdot \mathrm{sgn}(y \cdot w_d)$

-> For each component $\tilde{x}_d$, the attacker can shift x_d by at most P in either direction. The optimal direction is given by $-\mathrm{sgn}(y \cdot w_d)$ which pushes x_d as far as possible in the direction that maximally decreases the objective.

-> with $\tilde{x}_d = x_d - P \cdot \mathrm{sgn}(y \cdot w_d)$; we are ensuring that each term in $\tilde{x}^T$ is reduced as much as possible; so we're minimizing the overall value of $y(w_0 + \tilde{x}^T w)$

The objective function we aim to minimize is given by

$$y(w_0 + \tilde{x}^T w)$$

- (The attacker wants to find the perturbed vector $\tilde{x}$ such that this expression becomes as negative as possible.

$$\tilde{x} = x - P \cdot sgn(y \cdot w) \Rightarrow \tilde{x}_d = x_d - P\, sgn(y \cdot w_d)$$

If we substitute $\tilde{x}^T$ into the objective function:

$$y(w_0 + \tilde{x}^T w) = y(w_0 + (x - P \cdot sgn(y \cdot w))^T w)$$

$$y(w_0 + x^T w - P \cdot (sgn(yw))^T w)$$

but $sgn(y w_d) \cdot w_d = |w_d| \Rightarrow y \cdot (w_0 + x^T w - P\|w\|_1)$

$\|w\|_1 = \sum_{d=1}^{D} |w_d|$ is the l_1-norm of w

$$y \cdot (w_0 + x^T w) - P \cdot y \cdot \|w\|_1$$

since y is either $+/-1 \Rightarrow \|y \cdot w\|_1 = \|w\|_1 \rightarrow$

$$\Rightarrow y(w_0 + x^T w) - P\|yw\|_1$$

About the constraints: $|\tilde{x}_d - x_d| \leq P$

$$\tilde{x}_d = x_d - P \cdot sgn(y w_d)$$

$$|\tilde{x}_d - x_d| = |P \cdot sgn(y w_d)| = P$$

$$|\tilde{x}_d - x_d| \leq P \checkmark$$

$P\|yw\|_1 \Rightarrow (\|yw\|_1 = |yw_1| + |yw|_2 + --- + |yw_D|)$

$f(x) = \frac{\sigma}{2} \|x\|_2^2$ is cvx [Note task 5] ↳ sketch some plot of this stuff)

↓

cvx

→ TASK 2 completed (MATLAB CODE)

→ TASK 1 (Theoretical)

OPTION 1) MATLAB PLOT

OPTION 2)

(Convex Function) A function $f: C \Rightarrow \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^m$ is called convex (or convex over C) if

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y) \text{ for any } x, y \in C, \lambda \in [0,1]$$

In case when no domain is specified, then we naturally assume that f is defined over the entire space $\mathbb{R}^m$.

$N=1$, $D=1$ we simplify the function $f_D(w_0, w)$ from equation (2)

↓ (one point) ↓ (one dimension)

$$\frac{1}{N} \sum_{m=1}^{N} 1R\_ (y_m C_{w_0,w}(x_m)) \Longrightarrow 1R\_(y_1 C_{w_0,w}(x_1))$$

where $C_{w_0,w}(x_1) = \text{sign}(w_0 + w x_1)$ and $1R\_ = \begin{cases} 1 & \text{if } w < 0 \\ 0 & \text{if } w > 0 \end{cases}$

let's see: if we put: $w_0^1 = 1$, $w_1 = 1$; $w_0^2 = -1$, $w_2 = -1$; $x_1 = 1$ (input point), $y_1 = 1$ } this choice is just for simplicity.

• (1,1)

↓ $C_{1,1}(x_1) = \text{sgn}(1 + 1 \cdot 1) = \text{sign}(2) = 1$

$f_D(1,1) = 1R\_(1 \cdot 1) = 1R\_(1) \rightarrow 0$ [$f_D(1,1) = 0$ the classifier does not commit errors

• For (-1,-1)

$C_{-1,-1}(x_1) = \text{sgn}(-1 - 1 \cdot 1) = \text{sign}(-2) = -1$

$f_D(-1,-1) = 1R\_(1 \cdot -1) = 1R\_(-1) = 1$ (in this case we have an error)

By assumptions $\lambda = \frac{1}{2}$ (punto medio tra i 2 valori)

$\lambda = 0 \rightarrow \quad f(y) = f(y)$ ✓

$\lambda = 1 \rightarrow \quad f(x) = f(x)$ ✓

$\lambda = \frac{1}{2} \rightarrow$

$$f_D(\lambda(w_0^1, w^1) + (1-\lambda)(w_0^2, w^2)) \leq \lambda f_D(w_0^1, w^1) + (1-\lambda) f_D(w_0^2, w^2)$$

$$f_D\left(\frac{1}{2}(1,1) + \left(1-\frac{1}{2}\right)(-1,-1)\right) \leq f_D \frac{1}{2}(1,1) \; f_D\left(1-\frac{1}{2}\right)(-1,-1)$$

$$f_D\left(\frac{1}{2}-\frac{1}{2}, \frac{1}{2}-\frac{1}{2}\right) = f_D(0,0) = \quad C_{0,0}(x_1) = sgn(0+0\cdot 1) = sgn(0) = 0$$

$f_D(0,0) = 0 \rightarrow$ no errors

$$\frac{1}{2}(1,1) + \frac{1}{2}(-1,-1) = (0,0)$$

$$f_D(0,0) \leq \lambda f_D(1,1) + (1-\lambda) f_D(-1,-1)$$

$$\underbrace{f_D(0,0)}_{\substack{\parallel \\ 0}} \leq \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1$$

$0 \leq \frac{1}{2}$ (For $\lambda = \frac{1}{2}$ the condition is not satisfied)

TASK3

show that the function g_D defined in (4) is convex for any N & D
(N.B not only for N=1 and D=1)

$$\underbrace{\frac{1}{N}\sum_{m=1}^{N} h(y_m(w_0 + x_m^T w))}_{g_D(w_0,w)}$$

$g_D: R \times R^D \to R$

- h: some function (probably convex)
- $y_m \in \{-1,1\}$ labels of the dataset
- $x_m \in \mathbb{R}^D$ (feature vectors of the dataset)
- w_0 is the bias term
- $w \in \mathbb{R}^D$ is the weigth vector of the classifier.

1) let's consider only $h(y_m(w_0 + x_m^T w))$

- Since y_m is a constant (1/-1) it does not effect the convexity of the function.
- $w_0 + x_m^T$ is an affine function of (w_0, w) as it is a linear combination of w_0 and w

AN important property is that if h convex and differentiable then also the composition of a convex function with an affine transformation is also convex.

if h is convex $\Rightarrow$ $h(y_m(w_0 + x_m^T w))$ is convex

2) The summation of convex functions " $\frac{1}{N}\sum$ " remain convex, and multiplying by a positive constant (such as 1/N) does not effect the convexity. (this is true check chapter 7.4 (theorem 7.16))

Thus the entire function $g_D(w_0, w)$ is a convex function in (w_0, w)

1) Let f be a convex function defined over a convex set $C \subseteq \mathbb{R}^m$ and let $a \geq 0$. Then $a \cdot f$ is a convex function over C

PROOF:

$g(x) = af(x)$ $\quad x, y \in C$ and $\lambda \in [0,1]$. then.

$$\begin{aligned}
g(\lambda x + (1-\lambda)y) &= af(\lambda x + (1-\lambda)y) \quad (\underline{\text{DEF}} \text{ of } g) \\
&\leq a\lambda f(x) + a(1-\lambda)f(y) \quad (\text{convex of } f) \\
&= \lambda g(x) + (1-\lambda)g(y) \quad (\underline{\text{DEF}} \text{ of } g)
\end{aligned}$$

b) Let $f_1, f_2, ..., f_p$ be convex functions over a convex set $C \subseteq \mathbb{R}^m$, then the sum function $f_1 + f_2 + ... + f_p$ (Σ) is convex over C

PROOF:

Let $x, y \in C$ and $\lambda \in [0,1]$. For each $i = 1, 2, ..., p$ since f_i is convex, we have $f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y)$

summing the latter inequality over $i = 1, 2, ..., K$ yields the inequality:

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda) g(y)$$

For all $x, y \in C$ and $\lambda \in [0,1]$, where $g = f_1 + f_2 + --- + f_p$. We have thus established that the sum function is convex.

core part: $h(w) = (1-w)_+ = \max(0, 1-w) \Rightarrow$ hinge loss function

Trivial: • $u \geq 1$ $h(w) = 0$ (constant, hence convex)

ex. $\max(0, 1-1) = 0$

$\max(0, 1-10) = \max(0, -9) ---> 0$

$$\underbrace{\bullet\ w < 1,\ h(w) = 1 - w \text{ (which is linear)} \Downarrow \text{convex}}$$

To show that g_0 is convex, we need to prove that for any (w_0^1, w^1) and (w_0^2, w^2), and for any $\lambda \in [0,1]$, the following inequality holds:

$$g_0(\lambda w_0^1 + (1-\lambda) w_0^2, \lambda w^1 + (1-\lambda) W^2) \leq \lambda g_0(w_0^1, w^1) + (1-\lambda)(w_0^2, w^2)$$

LHS:

$$g_0(\lambda w_0^1 + (1-\lambda) w_0^2, \lambda w^1 + (1-\lambda) w^2) = \frac{1}{N} \sum_{m=1}^{N} h(y_m(\lambda(w_0^1 + x_m^T w^1) + (1-\lambda)(w_0^2 + x_m^T w^2)))$$

RHS:

$$\lambda g_0(w_0^1, w^1) + (1-\lambda) g_0(w_0^2, w^2) = \lambda \frac{1}{N} \sum_{m=1}^{N} h(y_m(w_0^1 + x_m^T w^1)) + (1-\lambda) \frac{1}{N} \sum_{m=1}^{N} h(y_m((w_0^2 + x_m^T w^2))$$

$h(w)$ is convex $\Rightarrow$ therefore it satisfies.

$$h(\lambda w_1 + (1-\lambda) w_2) \leq \lambda h(w_1) + (1-\lambda) h(w_2)$$

for any $w_1, w_2 \in \mathbb{R}$ and $\lambda \in [0,1]$

$$h(y_m(\lambda(w_0^1 + x_m^T w^1) + (1-\lambda)(w_0^2 + x_m^T w^2))) \leq \lambda h(y_m(w_0^1 + x_m^T w^1)) + (1-\lambda) h(y_m(w_0^2 + x_m^T w^2))$$

TASK 4

let's analyze also $\underbrace{\underbrace{\frac{1}{N}\sum_{m=1}^{N} h(y_m(w_0 + x_m^T w))}_{g_D(w_0,w)} + \underbrace{\rho\|w\|_2^2}_{r(w_0,w)}}_{g(w_0,w)}$ SQUARED EUCLIDEAN NORM

$\rho > 0$ $\|\cdot\|_2$ Euclidean Norm ($\|w\|_2 = \sqrt{w^T w}$)
$\hookrightarrow$ overfitting parameter (hyperparameter) to avoid overfitting

r: regularization term

$$\begin{cases} g(w_0,w) \text{ convex for any } N \text{ and } D \\ g(w_0,w) = g_D(w_0,w) + r(w_0,w) \\ r(w_0,w) = \rho\|w\|_2^2 \end{cases}$$

IF $\rho\|w\|_2^2$ is convex then:

- The function $g(w_0,w) = g_D(w_0,w) + \rho\|w\|_2^2$ is the sum of two convex functions:
  1. $g_D(w_0,w)$, which we already proved is convex
  2. $\rho\|w\|_2^2$ is convex for hypothesis.

  - The sum of two convex functions is still convex as we seen in TASK 3. (Anirbech. 7.4 (7.18 theorem))

$p > 0$ so we do not care.

$\|W\|_2 = \sqrt{w_1^2 + w_2^2 + \dots + w_D^2}$

$\hookrightarrow \sqrt{w^T w}$

$\|u\|_2^2 = u^T u$ or $\overset{\|w\|_2^2}{w^T w}$ (According to the notation we want to use)

] this is a quadratic function in w, and quadratic functions are convex as long as the associated matrix is positive semidefinite (Hessian Matrix)

Mathematically, we can prove convexity by checking the second derivative (or Hessian matrix) of $\|w\|_2^2$

$\nabla_w \|w\|_2^2 = 2w$

$H(w)$ is the matrix of second-order partial derivatives. Since $\|w\|_2^2$ is a quad function.

$H(w) = 2I$

$\hookrightarrow D \times D$