Supervised Learning, HW01

Ofer lipman, 201510435 and Daniel Shterenberg, 305199507

April 2021

1.1.1

Let $x, y \in \mathbb{R}^d_{\geq 0}$
Now lets take $z = \alpha x + (1-\alpha)y$
We notice that for each $z_i = \alpha x_i + (1-\alpha)y_i$
Since $x_i, y_i \geq 0$ and $0 \leq \alpha < 1$, then $\alpha x_i > 0$ and $(1-\alpha)y_i > 0$.
because adding 2 real non negative numbers will result in a real non negative number, then for each $z_i, z_i \in \mathbb{R}^d_{\geq 0}$ ■

1.1.2

We will prove the statement using an induction:

base case:
let n = 1:
then x (a set with 1 element), and according to the constrains, $\alpha = 1$.
and we get that the statement is true.

inductive step:
we will assume that if the statement is correct for each k = N.
such that for each $C \subseteq R^d$ such that $\{x_i \in C\}_{i=1}^{k}$, then $\Sigma_{i=1}^{k} \alpha_i x_i \in C$.
now lets prove for k + 1:
lets assume that there is a group $C \subseteq \mathbb{R}^d$
lets take $\{x_i \in C\}_{i=1}^{k+1}$
For

$$x = \Sigma_{i=1}^{k+1} \alpha_i x_i = \alpha_1 x_1 + \alpha_2 x_2 ... \alpha_{k+1} x_{k+1} = (\alpha_1 + ... + \alpha_k)(\frac{\alpha_1}{\alpha_1 + ... + \alpha_k} x_1 + ... \frac{\alpha_k}{\alpha_1 + ... + \alpha_k} x_k) + \alpha_{k+1} x_{k+1}$$

Since $\Sigma_{i=1}^{k} \frac{\alpha_i}{\alpha_1 + ... + \alpha_k} = 1$,
then by the induction hypothesis, $x' = \frac{\alpha_1}{\alpha_1 + ... + \alpha_k} x_1 + ... \frac{\alpha_k}{\alpha_1 + ... + \alpha_k} x_k \in C$
So $x = (\alpha_1 + ... + \alpha_k)x' + \alpha k + 1 x_{k+1}$.
Because C is a convex set, and $x', x_{k+1} \in C$, then $x \in C$ ■

1.1.3

We will disprove the statement:
lets take group (1,1),(2,2),(3,3),...,(10,10). this is a convex set, but the point (100,0) can not be represented by this convex combination ■

1.1.4

let $f : \mathbb{R} \to \mathbb{R}$ be such that $\forall x, y \in \mathbb{R}$: $f(y) \geq f(x) + f'(x)(y-x)$. We will prove that f is a convex function. Lets denote $z := \alpha \cdot x + (1-\alpha) \cdot y$. From the definition of f; $f(x) \geq f(z) + f'(z)(x-z)$ and $f(y) \geq f(z) + f'(z)(y-z)$ Now multiplying the first en-equality by α and the second one by $1-\alpha$ and adding both of them, we will get:

$$\begin{aligned} \alpha \cdot f(x) + (1-\alpha) \cdot f(y) &\geq f(z) + f'(z)(\alpha x + (1-\alpha)y - z) \\ &= f(z) \\ &= f(\alpha x + (1-\alpha)y) \end{aligned}$$

Therefore

$$f(\alpha x + (1-\alpha)y) \leq f(x) + (1-\alpha) \cdot f(y)$$

which means that f is a convex function from the definition. ■

1.2

Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $x_0 \in \mathbb{R}^d$.

1.2.1

$\forall h \in \mathbb{R}^d : \nabla f(x_0)[h] = \langle g_0, h \rangle$. We would prove that $g(x_0) = \nabla f(x_0)$.
Because the statement is true $\forall h \in \mathbb{R}^d$, then its also true for each one of the vectors e_i in the standard base vector base of $\mathbb{R}^d$, where

$$e_1 = (1, 0, .., 0), e_2 = (0, 1, 0, .., 0), ..., e_d = (0, ..., 0, 1)$$

Now, we will notice that $f(x_0)[e_i] = \frac{\partial f}{\partial x_i}(x_0)$.
On the other hand, let $g = (g_1, g_2, .., g_d)$. Therefore, $\langle g, e_i \rangle = g_i$ and from both of these statements we get that $g_i = \frac{\partial f}{\partial x_i}$ and that $g(x_0) = \nabla f(x_0)$. ■

1.2.2

let $h \in \mathbb{R}^{d_1}$ and let $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be a linear function. Lets examine the directional derivative of in direction h;

$$\nabla f(x)[h] = \lim_{t \to 0} \frac{f(x + th) - f(x)}{t} = \lim_{t \to 0} \frac{f(x) + tf(h) - f(x)}{t} = f(h)$$

The second equality follows from the linearity property. ■

1.2.3

1

let $f(x) = x^T Ax$.

$$\begin{aligned}
\nabla f(x)[h] = \lim_{t\to 0} \frac{f(x+th) - f(x)}{t} &= \lim_{t\to 0} \frac{(x+th)^T A(x+th) - x^T Ax}{t} \\
&= \lim_{t\to 0} \frac{x^T Ax + tx^T Ah + th^T Ax + t^2 x^T Ax - x^T Ax}{t} \\
&= x^T Ah + h^T Ax \\
&=^* x^T Ah + x^T A^T h \\
&= x^T (A + A^T)h \\
&= \langle (A + A^T)x, h\rangle
\end{aligned}$$

Therefore

$$\nabla f(x)[h] = \langle (A + A^T)x, h\rangle$$

$$\implies \nabla f(x) = (A + A^T)x$$

* $h^T Ax$ is a scalar and therefore $h^T Ax = (h^T Ax)^T = x^T A^T h$ ■

2

Let $f(X) = Tr\{X^T AX\}$.
We will recall that $Tr\{A^T B\} = \langle A, B\rangle$, and therefore $f(X) = \langle X, AX\rangle$.
Now, if we will denote $g(X) = X, h(X) = AX$, then $f(X) = \langle g(X), h(X)\rangle$ and from the product rule we will get that

$$\begin{aligned}
\nabla f(X)[H] = \nabla\langle g, h\rangle(X)[H] &= \langle \nabla g(X)[H], l(X)\rangle + \langle g(X), \nabla h()[H]\rangle \\
&= \langle H, AX\rangle + \langle X, AH\rangle \\
&= \langle (A + A^T)X, H\rangle
\end{aligned}$$

$$\implies \nabla f(X) = (A + A^T)X$$

■

3

Let $f(x) = \|y - Ax\|_2^2 = \langle y - Ax, y - Ax\rangle$
Following the product rule,

$$\begin{aligned}
\nabla f(x)[h] &= \langle -Ah, y - Ax\rangle + \langle y - Ax, -Ah\rangle \\
&= 2\langle y - Ax, -Ah\rangle \\
&= -2\langle A^T (y - Ax), h\rangle
\end{aligned}$$

$$\implies \nabla f(x) = -2A^T (y - Ax)$$

■

4

Let $f(X) = \|Y - AX\|_F^2 = \langle Y - AX, Y - AX \rangle$
Following the product rule,

$$\nabla f(X)[H] = \langle -AH, Y - AX \rangle + \langle Y - AX, -AH \rangle = -2\langle A^T(Y - AX), H \rangle$$

$$\Longrightarrow \nabla f(X) = -2A^T(Y - AX)$$

$\blacksquare$

5

Let $f(X) = \langle X^T A, Y^T \rangle$. We will denote $g(X) = X^T A, h(X) = Y$.

$$\nabla g(X) = H^T A, \nabla h(X) = 0$$

Following the product rule,

$$\nabla f(X)[H] = \langle H^T A, Y^T \rangle + \langle X^T A, 0 \rangle = \langle H^T A, Y^T \rangle = \langle A, HY^T \rangle = \langle AY, H \rangle$$

$$\Longrightarrow \nabla f(X) = AY$$

$\blacksquare$

6

Let $f(x) = a^T g(x)$. In order to find $\nabla f(x)$ we will need to find $\nabla g(x)$.

$$\begin{aligned}
\nabla g(x)[h] &= \lim_{t \to 0} \frac{g(x + th) - g(x)}{t} \\
&= \lim_{t \to 0} \frac{1}{t} \cdot \left(\begin{pmatrix} g(x_1 + th_1) \\ g(x_2 + th_2) \\ \dots \\ g(x_d + th_d) \end{pmatrix} - \begin{pmatrix} g(x_1) \\ g(x_2) \\ \dots \\ g(x_d) \end{pmatrix} \right) \\
&= \lim_{t \to 0} \frac{1}{t} \cdot \begin{pmatrix} g(x_1 + th_1) - g(x_1) \\ g(x_2 + th_2) - g(x_2) \\ \dots \\ g(x_d + th_2) - g(x_d) \end{pmatrix} \\
&= \begin{pmatrix} g'(x_1)h_1 \\ g'(x_2)h_2 \\ \dots \\ g'(x_d)h_d \end{pmatrix} = g'(x) \circ h = diag(g'(x)) \cdot h.
\end{aligned}$$

$$\nabla f(x)[h] = \langle a, \nabla g(x)[h] \rangle = \langle a, diag(g'(x)) \cdot h \rangle = \langle diag(g'(x)) \cdot a, h \rangle$$

$$\Longrightarrow \nabla f(x) = diag(g'(x)) \cdot a$$

$\blacksquare$

7

Let $f(X) = \langle A, \log[X] \rangle$. We will denote $g(X) = \log[X]$.

$$\nabla g(X)[H] = \lim_{t \to 0} \frac{g(X + tH) - g(X))}{t} = \lim_{t \to 0} \frac{log[X + tH] - log[X]}{t} = \lim_{t \to 0} \frac{log[\frac{X+tH}{X}]}{t}$$

If we will denote with M the matrix $log[\frac{X+tH}{X}]$ where $M_{i,j} = \log[1 + t \cdot \frac{H_{i,j}}{X_{i,j}}]$ then

$$\lim_{t \to 0} M = \lim_{t \to 0} \frac{log[\frac{X+tH}{X}]}{t} = H \circ X^{\circ -1}$$

Following the product rule:

$$\nabla f(X)[H] = \langle 0, log[X] \rangle + \langle A, H \circ X^{\circ -1} \rangle = \langle A \circ X^{\circ -1}, H \rangle$$

$$\implies \nabla f(X) = A \circ X^{\circ -1}$$

$\blacksquare$

8

Let $f(X) = \langle a, diag(X) \rangle$. We will notice that $a = (a_1, a_2, ..., a_d)$ and that $diag(a)$ is the matrix $A : A_{ii} = a_i, A_{ij} = 0$ where $i \neq j$ and from that we can derive that $f(X) = \langle a, diag(X) \rangle = \langle diag(a), X \rangle$

$$\nabla f(X)[H] = \langle diag(a), H \rangle$$

$$\implies \nabla f(X) = diag(a)$$

$\blacksquare$

1.3

In the Python notebook

1.4

1.4.1

Let $G(x, y) = sin(x + y)$

1

$\min_x \max_y G(x, y) =?$
we will notice that for some const x_0, $\max_y G(x_0, y) = \max_y sin(x_0 + y) = 1$ and therefore, $\min_x \max_y G(x, y) = \min_x \max_y sin(x + y) = \min_x 1 = 1$ $\blacksquare$

2

$\max_y \min_x G(x, y) =?$
we will notice that for some const y_0, $\min_x G(x, y_0) = \min_x sin(x + y_0) = -1$ and therefore, $\max_y \min_x G(x, y) = \max_y \min_x sin(x + y) = \max_y -1 = -1$ ■

1.4.2

$f(x) = \frac{x^T Ax}{x^T x} = \frac{\langle x, Ax\rangle}{\langle x, x\rangle}$ where $A \in \mathbb{R}^{d\times d}$ and symmetric.

a.

let a be a non-zero scalar. We will notice that

$$f(a \cdot x) = \frac{\langle a \cdot x, a \cdot Ax\rangle}{\langle a \cdot x, a \cdot x\rangle} = \frac{a^2 \cdot \langle x, Ax\rangle}{a^2 \cdot \langle x, x\rangle} = \frac{\langle x, Ax\rangle}{\langle x, x\rangle} = f(x)$$

Therefore, without loss of generality, we can assume that $\|x\|_2^2 = 1$.
For such x,

$$f(x) = \frac{\langle x, Ax\rangle}{\|1\|_2^2} = \langle x, Ax\rangle$$

and

$$\min_x f(x) = \min_x \langle x, Ax\rangle = \min_x x^T Ax$$

which is excatly what we wanted to prove.

b.

We are trying to minimize $f(x)$ under the constraint $x^T x = 1$ i.e

$$\min_x x^T Ax$$
$$s.t\ x^T x = 1$$

We will denote as $g(x) = x^T x - 1$ and then we would be able to write the Lagrangian of the constraint as

$$\mathcal{L}(x, \lambda) = f(x) - \lambda \cdot g(x) = x^T Ax - \lambda \cdot (x^T x - 1)$$

c.

We will take the derivative of $\mathcal{L}(x, \lambda)$ with respect to x.

$$\begin{aligned}\nabla_x \mathcal{L}(x, \lambda) &= \frac{\partial \mathcal{L}(x, \lambda)}{\partial x} \\ &= \frac{\partial f(x) - \lambda \cdot g(x))}{\partial x} \\ &= \frac{\partial \langle x, Ax\rangle}{\partial x} - \lambda \frac{\partial (x^T x - 1)}{\partial x} \\ &= (A + A^T) \cdot x + 2\lambda \cdot x.\end{aligned}$$

Now lets compare this expression to 0.

$$\nabla_x \mathcal{L}(x, \lambda) = 0 \iff (A + A^T) \cdot x - 2\lambda \cdot x = 0 \iff$$

$$2A \cdot x = 2\lambda \cdot x \iff A \cdot x = \lambda \cdot x$$

Where the transition between the 2nd statement to the 3rd statement is due to the symmetry of A. ■

2.1

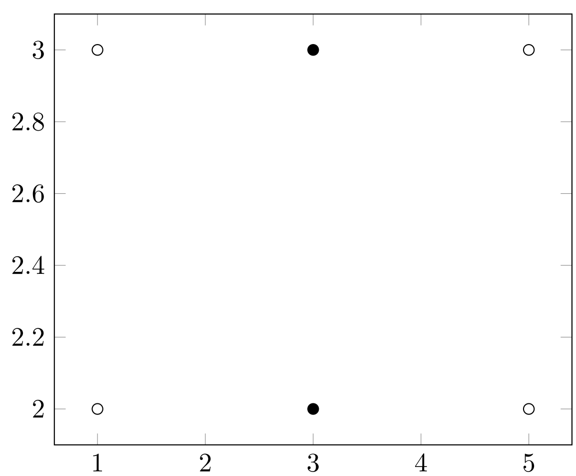
2.1.1

1

2

2.1.2

We will disprove with the following example:
lets take the following points (1,2), (1,3), (5,2), (5,3):
Now, if the centroids starting positions are μ_1 =(3, 2) and μ_1 =(3,3), then the clusters are: $D_1 = \{(1,2),(5,2)\}$ and $D_2 = \{(1,3),(5,3)\}$



The centroids are the center of the cluster, and they will no change.
but the global solution is μ_1 =(1, 2.5) and μ_1 =(5,2.5)