Introduction to Machine Learning Course

Short HW4 – Optimization, Regression, and Boosting

Submitted <u>individually</u> by Thursday, 22.06.23, at 23:59.

You may answer in Hebrew or English and write on a computer or by hand (but be clear).

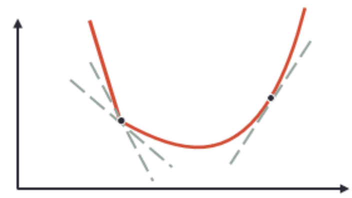
Please submit a PDF file named like your ID number, e.g., 123456789.pdf.

Bonus (maximal grade is 100): Writing on a computer (using LyX/LaTeX, Word + Equation tool, etc.) = 2 pts.

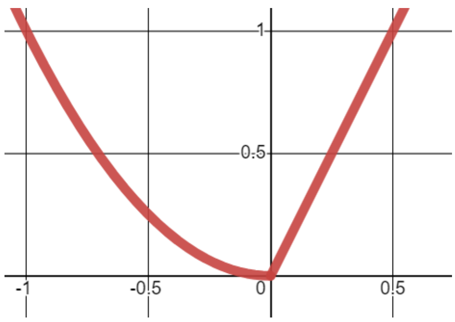
Part A – Optimization

As we saw in Tutorial 08, subgradients generalize gradients to convex functions which are not necessarily differentiable. Notice: you can solve this exercise even before watching Tutorial 08.

Definition: the set of subgradients of $f: V \to \mathbb{R}$ at point $\boldsymbol{u} \in V$ is:



$$\partial f(\boldsymbol{u}) \triangleq \{\boldsymbol{q} \in V \mid \forall \boldsymbol{v} \in V: f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boldsymbol{q}^\top(\boldsymbol{v} - \boldsymbol{u})\}.$$



1. Let $f(x) = \begin{cases} x^2, & x < 0 \\ 2x, & x \geq 0 \end{cases}$.
   1.1. Is f convex? No need to explain.
   1.2. Propose a sub-derivative function g for f. That is, $g \in \partial f$. Use the above definition to prove that $g(u) \in \partial f(u), \forall u \in \mathbb{R}$.
   1.3. Set a learning rate of $\eta = 0.25$ and a starting point $x_0 = -1.5$. Running subgradient descent, will the algorithm converge to a minimum? Prove your answer by filling the following table like we did in Tutorial 07 using as many rows as needed.

i	$\boldsymbol{x_i}$	$\boldsymbol{f(x_i)}$	$\boldsymbol{\frac{\partial}{\partial x} f(x_i) = g(x_i)}$
0	-1	1	
1			
⋮			

   1.4. Repeat 1.3 with $\eta = 1,\ x_0 = -1.5$.

Part B – Regression

2. Consider the ridge regression problem: $\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\left(\frac{1}{m}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2\right)$.

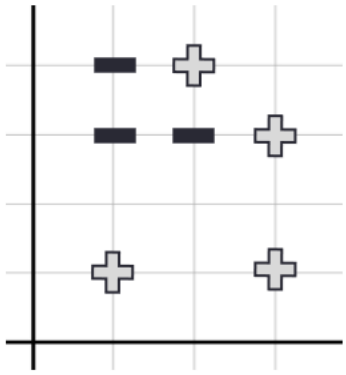
   Also denote the singular value decomposition (SVD) of $\mathbf{X}$ as $\underbrace{\mathbf{X}}_{m\times d} = \underbrace{\mathbf{U}}_{m\times m}\underbrace{\boldsymbol{\Sigma}}_{m\times d}\underbrace{\mathbf{V}^\top}_{d\times d}$, where $\mathbf{U},\mathbf{V}$ are real orthonormal matrices.

   a. Prove that the matrix $(\mathbf{X}^\top\mathbf{X} + m\lambda\mathbf{I})$ is positive definite.

   b. Prove that the closed form solution is $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + m\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{y}$ and briefly explain why it is unique.

   c. Prove that additionally, $\hat{\mathbf{w}} = \mathbf{V}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} + m\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\boldsymbol{y}$.

Part C – Boosting



3. Given the following data with binary labels ("+", "-").

   We run AdaBoost with Decision stumps as weak classifiers.
   The sizes of the shapes in the figures indicate the probabilities that the algorithm assigns to each sample (high probability = large shape). Initially, the algorithm starts from a uniform distribution.

   Only some of the following figures depict possible distributions that can be obtained after <u>one</u> iteration of AdaBoost. **Which ones?** For each such distribution, propose a weak classifier that can lead to its figure (use a <u>clear</u> drawing or a short description of that classifier).

(a) **(b)** **(c)** **(d)** **(e)**