Introduction to Machine Learning Course

Dry 5 – Regression and Boosting

Submitted individually by Sunday, 20.06, at 23:59. Each day of delay costs 5 points.

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.) = 3 pts.

Part A – Regression

Consider the least squares problem with a matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ and a vector $\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{m}}$: $\quad \operatorname{argmin}_{\boldsymbol{w}} \underbrace{\|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2}_{\mathcal{L}(\boldsymbol{w})}$.

Remember that $\mathcal{L}(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = \sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$.

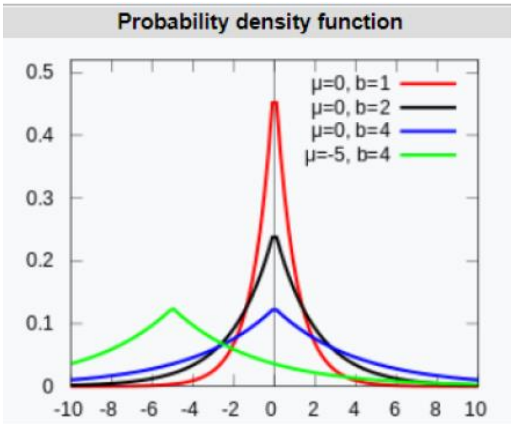
1. We now prove that the loss is convex in $\boldsymbol{w}$.
   1.1. Derivate the first order derivative $\frac{\partial}{\partial w_k}\mathcal{L}(\boldsymbol{w})$.
   1.2. Derivate the second order derivative $\frac{\partial^2}{\partial w_j \partial w_k}\mathcal{L}(\boldsymbol{w})$.
   1.3. Conclude that the Hessian is $\nabla^2_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}) = 2\mathbf{X}^\top\mathbf{X}$.
   1.4. Prove that $\mathcal{L}$ is convex in $\boldsymbol{w}$.

2. Consider a noisy linear model where $y = \langle \boldsymbol{w}, x\rangle + \varepsilon$, for:
   - Given examples $\boldsymbol{x} \in \mathbb{R}^d$
   - An unknown weight vector $\boldsymbol{w} \in \mathbb{R}^d$
   - Random i.i.d noise ε

   In Lecture 09 (slides 09-13), we showed that when $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, the solution of the least squares formulation is a Maximum-Likelihood Estimator (MLE) of the unknown $\boldsymbol{w}$.

   Prove that when $\varepsilon_i \sim \text{Laplace}(0, b)$, the MLE for $\boldsymbol{w}$ corresponds to the solution of the least absolute deviation problem:

$$\operatorname{argmin}_{\boldsymbol{w}} \underbrace{\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{w}^\top \boldsymbol{x}_i - y_i|}_{\mathcal{L}_{\text{abs}}(\boldsymbol{w})}.$$



   That is, prove that:

$$\underbrace{\operatorname{argmax}_{\boldsymbol{w}} \prod_{i=1}^{m} P(y_i, \boldsymbol{x}_i; \boldsymbol{w})}_{\text{Maximum-Likelihood Estimator}} = \underbrace{\operatorname{argmin}_{\boldsymbol{w}} \frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{w}^\top \boldsymbol{x}_i - y_i|}_{\text{Least absolute deviation}}.$$

   Reminder: The Laplacian pdf's is $p(w_j | \mu, b) = \frac{1}{2b}\exp\left\{-\frac{|w_j - \mu|}{b}\right\}$.

   Its statistics are $\mathbb{E}[w_j] = \mu$ and $\text{Var}[w_j] = 2b^2$.

Part B – Boosting

3. Prove that when running AdaBoost, the distribution is updated such that the error of the chosen weak classifier h_t, w.r.t the updated distribution $D_i^{(t+1)}$, is exactly $\frac{1}{2}$.

   That is, prove that $\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i} = \frac{1}{2}$.

   Hint: You can fill the missing steps in the following derivation:

$$\sum_i D_i^{(t+1)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i} = \cdots = \frac{\epsilon_t}{\epsilon_t + (1-\epsilon_t)\exp\{-2\alpha_t\}} = \cdots = \frac{1}{2}.$$