Problem 1 [10 marks]. We mentioned that directly applying linear regression to classification labels $\{0, 1\}$ is not a good idea. Explain the reason [5 marks]. Does tuning regularization help (e.g., increasing or decreasing the coefficient of ℓ_2-penalty)? Why or why not? [5 marks].

> **Note:** One or a few sentences suffice for each question. Long answers with wrong or not understandable statements will result in mark deduction.

Problem 2 [30 marks]. Consider a logistic regression model $y = \sigma(\mathbf{w}^\top \mathbf{x} + b)$ and a two-way classification model $\mathbf{y} = \mathrm{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})$ for d-dimensional input $\mathbf{x} \in \mathbb{R}^d$.

a) [10 marks] Write out the formulas of the sigmoid and softmax functions.

b) [10 marks] How many model parameters do we have for the logistic regression model and the softmax regression model, respectively?

c) [10 marks] Given the same set of training data, which model (logistic vs softmax) is more likely to overfit? And why?

Hint: A d-dimensional vector counts d parameters. No derivation or proof is needed.

Problem 3 [30 marks]. A sample has d features, $\mathbf{x} = (x_1, \cdots, x_d)^\top \in \mathbb{R}^d$; the target is a real number $t \in \mathbb{R}$. We would like to consider point-wise quadratic features $x_1^2, \cdots, x_d^2$ in addition to the original ones. In other words, the augmented features will be $\widetilde{\mathbf{x}} = (x_1, \cdots, x_d, x_1^2, \cdots, x_d^2, 1)^\top \in \mathbb{R}^{2d+1}$. We denote the regression model by $h(\mathrm{x}) = \widetilde{\mathbf{w}}^\top \widetilde{\mathbf{x}}$.

a) [5 marks] Give the mean square error loss J_{MSE} on the training set $\mathcal{D} = \{(\mathbf{x}^{(m)}, t^{(m)})\}_{m=1}^M$.

b) [10 marks] What's the probabilistic interpretation for this MSE? (5 marks for what variables coming from what distributions, 5 marks for the generic formulation of the parameter estimation criterion). Proof is not required.

c) [10 marks] Compute $\frac{\partial}{\partial \widetilde{w}_i} J_{\mathrm{MSE}}$, where $\widetilde{w}_i$ is an element in $\widetilde{\mathbf{w}}$ for $i = 1, \cdots, 2d+1$. Give a few derivation steps.

d) [5 marks] We observe the learned model is underfitting, leading to low performance. Is collecting more data a good approach to improve performance? Why or why not?

Hint: The constant in MSE does not matter. However, it must be consistent in a) and c).

Problem 4 [30 marks]. One idea of using a linear function $y = \mathbf{w}^\top \mathbf{x}$ for classification is to apply the max-margin loss. Suppose the target label is $t \in \{-1, 1\}$, the max-margin loss for a sample is defined to be $J^{(m)} = \max\{0, 1 - t^{(m)} \cdot y^{(m)}\}$. Here, the function $\max\{a, b\}$ chooses the maximum value, for example, $\max\{0, 0.3\} = 0.3$, $\max\{0, -0.2\} = 0$.

a) [10 marks] Draw two curves to show how $J^{(m)}$ responds according to $y^{(m)}$, for $t^{(m)} = -1$ and $t^{(m)} = 1$, respectively.

b) [10 marks] Prove that $J^{(m)}$ is convex in $\mathbf{w}$. *Hint*: $\max$ is not a differentiable function.

c) [10 marks] Give an algorithm for solving this optimization problem. If you give a closed-form solution, derive the formula. If you give a gradient-based approach, write the pseudo-code (similar to lecture notes) and compute the gradient.

Hint: Useful identity: $\frac{\partial}{\partial \mathbf{u}} \mathbf{u}^\top \mathbf{v} = \mathbf{v}$

End of the exam

Scrap paper

- Additional pages are available upon request.
- May be used as an answer sheet if you mark problem numbers clearly.