Time: 2 hours

Exercise 1 [8 points; max 2 pages.]

Consider the *binary* classification problem with 0-1 loss.

1. Provide a formal definition of the problem, describing the data, the learner's input, the learner's output, the loss function, the assumed generative model for the data, the learner's goal, and what choices the learner has to make when trying to solve the problem.

2. Assume the hypothesis class is $\mathcal{H}$, the data generative distribution is $\mathcal{D}$, and the training data is S. Provide the definition of the training error $L_S(h)$ and generalization error $L_{\mathcal{D}}(h)$, $h \in \mathcal{H}$, and prove that the expectation (over the distribution of the training set) of $L_S(h)$ is equal to $L_{\mathcal{D}}(h)$.

3. Use the result above to argue that the ERM procedure can be appropriate when the training data is large enough.

4. Provide the definition of ε-representative sample, and prove that if the training set S is $\frac{\varepsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$), then any output h_S of the ERM procedure (using $\mathcal{H}$) satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

   (Provide a motivation for each step of the proof.)

Exercise 2 [8 points; max 2 pages.]

Consider the regression problem with training data $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, m$. Assume the hypothesis class is given by linear models, the squared loss is used, and Tikhonov regularization is used.

1. Formally define the hypothesis class, the loss function, and the function that is optimized by ERM with Tikhonov regularization (i.e., the objective function).

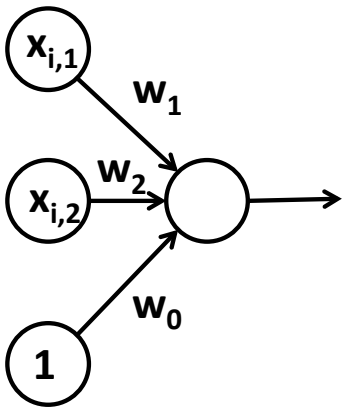
2. Provide the formula for the best model (i.e., which optimizes the objective function) learned from data in the context above. (The derivation is not required.)

3. In Tikhonov regularization, the parameter λ regulates the tradeoff between the empirical risk and the complexity of the model. Describe one approach to choose λ.

Exercise 3 [8 points; max 2 pages.]

Consider the regression problem, with training data $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i = [x_{i,1}, x_{i,2}] \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$ for $i = 1, \dots, m$.

Assume the hypothesis class $\mathcal{H}$ is given by the simple neural network with the architecture described in the figure below, where w_0, w_1, w_2 are the edges' weights.



Assume the activation function for the output node is $\sigma(z) = e^z$ (the final prediction is the output of the output node), and the loss function is $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^4$.

1. Derive a closed-form expression for the hypotheses in $\mathcal{H}$.
2. Describe the stochastic gradient descent (SGD) algorithm (in general). What is the main advantage of SGD with respect to the gradient descent algorithm?
3. Write the SGD update for learning a model from the hypothesis class $\mathcal{H}$ above.

Exercise 4 [8 points; max 2 pages.]

Consider the clustering problem.

1. Briefly describe *linkage-based clustering*, what is the input, what is the output, the general algorithm it employs, and three common termination conditions.

2. Consider *single linkage* clustering. Describe how it is obtained by the general *linkage-based clustering* (no pseudocode needed).

3. Show the output of single linkage clustering when the input is given by the points in $\mathbb{R}^2$ shown as crosses below and the termination condition is given by having the points partitioned in $k = 2$ clusters. Briefly describe how the algorithm reaches such output.