IV. Maximum Likelihood Estimation
V. Perceptron
VI. Regularization
VII. General ML concepts (generalization)

Consider a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ and a probabilistic model $\mathbb{P}(y_i \mid \boldsymbol{x}_i; \boldsymbol{w})$ which specifies for each i the probability of seeing outcome y_i given feature $\boldsymbol{x}_i$ and parameter $\boldsymbol{w}$. Which of the following is the Maximum Likelihood Estimation (MLE) for $\boldsymbol{w}$?

(A) $\operatorname{argmax}_{\boldsymbol{w}} \sum_{i=1}^{n} \mathbb{P}(y_i \mid \boldsymbol{x}_i; \boldsymbol{w})$

(B) $\operatorname{argmax}_{\boldsymbol{w}} \Pi_{i=1}^{n} \mathbb{P}(y_i \mid \boldsymbol{x}_i; \boldsymbol{w})$

(C) $\operatorname{argmax}_{\boldsymbol{w}} \sum_{i=1}^{n} \ln \mathbb{P}(y_i \mid \boldsymbol{x}_i; \boldsymbol{w})$

(D) $\operatorname{argmax}_{\boldsymbol{w}} \Pi_{i=1}^{n} \ln \mathbb{P}(y_i \mid \boldsymbol{x}_i; \boldsymbol{w})$

Which of the following statements are true about Maximum likelihood estimation (MLE)?

(A) To do MLE of some parameter $\boldsymbol{w}$, we need to first write a probablistic model $\mathbb{P}[y|\boldsymbol{x};\boldsymbol{w}]$ which specifies how the label y of a datapoint $\boldsymbol{x}$ is generated based on $\boldsymbol{w}$.
(B) If we have a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$, where each outcome y_i is generated by a probabilistic model $\mathbb{P}[y_i = 1|\boldsymbol{x}_i;\boldsymbol{w}] = \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)$ where $\sigma(\cdot)$ is the sigmoid function, then MLE for $\boldsymbol{w}$ is equivalent to empirical risk minimization on the logistic loss.
(C) If we have a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, where each outcome y_i is generated by a probabilistic model $y_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + \epsilon_i$ with ϵ_i being an independent Gaussian noise with zero-mean, then MLE for $\boldsymbol{w}$ is equivalent to least squares estimation.
(D) Maximum a posteriori probability (MAP) estimation is an extension of MLE where we assume a prior over $\boldsymbol{w}$.

IV. Maximum Likelihood Estimation
V. Perceptron
VI. Regularization
VII. General ML concepts (generalization)

We will investigate the perceptron algorithm in this question (the algorithm is reproduced in Algorithm 2). The perceptron algorithm gets access to a dataset of n instances $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. It outputs a linear classifier $y = \text{SIGN}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$. Assume $\boldsymbol{x}_i \neq \mathbf{0}\ \forall\ i \in \{1, \ldots, n\}$.

Algorithm 2: Perceptron

Input: A training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, number of iterations T
Initialize $\boldsymbol{w}_{(0)} \leftarrow \mathbf{0}$;
for t *in* $\{0, \ldots, T-1\}$ **do**
  Pick a data point $(\boldsymbol{x}_i, y_i)$ randomly
  Make a prediction $\hat{y} = \text{SIGN}({\boldsymbol{w}_{(t)}}^T \boldsymbol{x}_i)$ using $\boldsymbol{w}_{(t)}$
  if $\hat{y} \neq y_i$ **then**
    $\boldsymbol{w}_{(t+1)} \leftarrow \boldsymbol{w}_{(t)} + y_i \boldsymbol{x}_i$
  else
    $\boldsymbol{w}_{(t+1)} \leftarrow \boldsymbol{w}_{(t)}$

As the algorithm proceeds, suppose the same weight vector is seen twice, despite at least one update in between. In particular, suppose there is some j and k where $j < k$ such that $\boldsymbol{w}_{(j)} = \boldsymbol{w}_{(k)}$, and there is at least one ℓ where $j < \ell < k$ such that $\boldsymbol{w}_{(j)} \neq \boldsymbol{w}_{(\ell)}$. We will show that if this happens, then the given dataset is not linearly separable. To prove this, follow the following two steps.

(a) Write down an expression which relates $\boldsymbol{w}_{(j)}$ and $\boldsymbol{w}_{(k)}$. Your expression will involve the datapoints observed in the intermediate iterations. (5 points).

(b) Now suppose there is some linear classifier $\boldsymbol{w}^*$ which classifiers all datapoints perfectly, i.e. $y_i = \text{SIGN}({\boldsymbol{w}^*}^{\mathrm{T}} \boldsymbol{x}_i)$ for all $i \in \{1, \ldots, n\}$. Use your expression from the previous part and the fact that $\boldsymbol{w}_{(j)} = \boldsymbol{w}_{(k)}$ to arrive at some contradiction, hence proving that the dataset cannot be linearly separable if $\boldsymbol{w}_{(j)} = \boldsymbol{w}_{(k)}$. (5 points).

IV. Maximum Likelihood Estimation
V. Perceptron
VI. Regularization
VII. General ML concepts (generalization)

Your first task is to fit a linear regression model for some customer data. The data has $d = 100$ real-valued attributes, and $n = 50$ datapoints are available.

(a) What could go wrong if you try to naively find the least-squares estimator on the data?

(b) How can you rectify the issue you identified in the previous question?

In class, we discussed that if we use Newton's method to solve the least square optimization problem then it only takes one step to converge. We will prove this statement in this problem. Let

$$F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Recall that Newton's method updates the parameters as follow:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \boldsymbol{H}_t^{-1}\nabla F(\boldsymbol{w}^{(t)})$$

where $\boldsymbol{H}_t = \nabla^2 F(\boldsymbol{w}^{(t)}) \in \mathbb{R}^{d\times d}$ is the Hessian matrix of the objective function evaluated at $\boldsymbol{w}^{(t)}$, i.e. for every index $u, v \in \{1, \ldots, d\}$ the (u, v)-th entry of $\boldsymbol{H}_t$ is $H_t(u, v) = \frac{\partial^2}{\partial w_u \partial w_v} F(\boldsymbol{w}) \mid_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$.

(1) Find the Hessian of the least square objective function $F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$.

In class, we discussed that if we use Newton's method to solve the least square optimization problem then it only takes one step to converge. We will prove this statement in this problem. Let

$$F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Recall that Newton's method updates the parameters as follow:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \boldsymbol{H}_t^{-1}\nabla F(\boldsymbol{w}^{(t)})$$

where $\boldsymbol{H}_t = \nabla^2 F(\boldsymbol{w}^{(t)}) \in \mathbb{R}^{d\times d}$ is the Hessian matrix of the objective function evaluated at $\boldsymbol{w}^{(t)}$, i.e. for every index $u, v \in \{1, \ldots, d\}$ the (u, v)-th entry of $\boldsymbol{H}_t$ is $H_t(u, v) = \frac{\partial^2}{\partial w_u \partial w_v} F(\boldsymbol{w}) \mid_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$.

(1) Find the Hessian of the least square objective function $F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$.

(2) What is the sufficient condition of λ to apply Newton's method? (Hint: $\boldsymbol{H}$ is invertible if it is positive definite, *i.e.*, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{u} > 0$ for $\forall \boldsymbol{u}$.)

In class, we discussed that if we use Newton's method to solve the least square optimization problem then it only takes one step to converge. We will prove this statement in this problem. Let

$$F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Recall that Newton's method updates the parameters as follow:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \boldsymbol{H}_t^{-1}\nabla F(\boldsymbol{w}^{(t)})$$

where $\boldsymbol{H}_t = \nabla^2 F(\boldsymbol{w}^{(t)}) \in \mathbb{R}^{d\times d}$ is the Hessian matrix of the objective function evaluated at $\boldsymbol{w}^{(t)}$, i.e. for every index $u, v \in \{1, \ldots, d\}$ the (u, v)-th entry of $\boldsymbol{H}_t$ is $H_t(u, v) = \frac{\partial^2}{\partial w_u \partial w_v} F(\boldsymbol{w}) \mid_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$.

(1) Find the Hessian of the least square objective function $F(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda||\boldsymbol{w}||_2^2$.

(2) What is the sufficient condition of λ to apply Newton's method? (Hint: $\boldsymbol{H}$ is invertible if it is positive definite, *i.e.*, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{u} > 0$ for $\forall \boldsymbol{u}$.)

(3) Show that given any initialization $\boldsymbol{w}^{(0)}$, if λ satisfies the condition in (2), after one iteration of Newton's method we obtain the optimal $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. (Here $\boldsymbol{X}$ is the $n \times d$ data matrix with one row per data point, and $\boldsymbol{y}$ is the n-vector of their labels.)

IV. Maximum Likelihood Estimation
V. Perceptron
VI. Regularization
VII. General ML concepts (generalization)

Which of the following are true statements about supervised learning?

(A) The test set should not be used to train the model, but can be used to tune hyper-parameters.
(B) The generalization gap (difference between test and training errors) generally decreases as the size of the training set increases.
(C) We cannot estimate the risk of a predictor (its average error on the data distribution) solely with the data used to train it.
(D) If training and test data are drawn from different distributions, then low error on the training set may not guarantee low error on the test set even if the size of the training set is sufficiently large.

Which of the following statements are true?

(A) In supervised learning, we assume that we are provided the desired output labels for each datapoint.
(B) A classifier that attains 100% accuracy on the training set and 70% accuracy on the test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on the test set.
(C) A model which has high training error and high test error is said to be underfitting.
(D) It is not advisable to use the test set to tune hyperparameters of our machine learning model.

Which of the following statements are true about generalization?

(A) A model which always makes the same prediction on any input datapoint will have small difference in accuracy between training and test data.
(B) The gap between training and test accuracies will generally *increase* as the size of the training dataset *increases*.
(C) To measure the expected risk of a machine learning model, we usually estimate the risk on new datapoints drawn i.i.d. from the data distribution.
(D) If the i.i.d. assumption is not valid, then our model may not perform well on new unseen examples when deployed in the real world.