Optimization and Machine Learning, Spring 2020
Homework 1

(Due Wednesday, Mar. 18 at 11:59pm (CST))

March 14, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variables and one continuous/binary response.
   (a) Please define the input and output variables, and show a linear relationship between them. (5 points)
   (b) Please define a data matrix and corresponding response vector, and find your i-th ($i = 1, ..., N$) sample with its response. (5 points)
   (c) Please use the least squares to estimate the parameters of the linear model in (a) based on the dataset in (b), and explain in which case the solution is unique. (10 points)
   (d) Is there any way to get an unique closed-form solution once the least squares fails? If yes, please show how do you obtain the solution. (5 points)
   (e) How can you select the best model in (d) based only on your training data. (5 points)

2. Given the input variables $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$, the Expected Prediction Error (EPE) is defined by
$$\text{EPE}(\hat{f}) = \mathbb{E}[L(Y, \hat{f}(X))], \tag{1}$$
where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{f}(X))$ is a loss function measuring the difference between the estimated $\hat{f}(X)$ and observed Y.
   (a) Given the squared error loss $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$, please derive the regression function $\hat{f}(x) = \mathbb{E}(Y|X = x)$ by minimizing $\text{EPE}(\hat{f})$ w.r.t. $\hat{f}$. (5 points)
   (b) Please explain why the nearest neighbors is an approximation to the regression function in (a). (5 points)
   (c) Please explain how the least squares approximates the regression function in (a). (5 points)
   (d) Please discuss the difference between the nearest neighbors and the least squares based on your results in (b) and (c). (5 points)

3. Given a set of observation pairs $(x_1, y_1) \cdots (x_N, y_N)$. By assuming the linear model is a reasonable approximation, we consider fitting the model via least squares approaches, in which we choose coefficients β to minimize the residual sum of squares (RSS),
$$\hat{\beta}_0,\ \hat{\beta} = \underset{\beta_0,\ \beta}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta x_i)^2.$$
   (a) Show that
$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned} \tag{2}$$
   where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the sample means. (3 points)
   (b) Using (2), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$. (2 points)

4. Given a set of training data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$ from which to estimate the parameters $\boldsymbol{\beta}$, where each $\boldsymbol{x}_i = [x_{i1}, \cdots, x_{ip}]^T$ denotes a vector of feature measurements for the ith sample. Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we aim at minimizing
$$\text{RSS}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2. \tag{3}$$

(a) Show that $\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T \boldsymbol{W} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})$ for an appropriate diagonal matrix $\boldsymbol{W}$, and where $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^T$ and $\boldsymbol{y} = [y_1, \cdots, y_N]^T$. State clearly what $\boldsymbol{W}$ is. (1 points)

(b) By finding the derivative $\nabla_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta})$ and setting that to zero, write the normal equations to this weighted setting and give the value of $\boldsymbol{\beta}$ that minimizes $\text{RSS}(\boldsymbol{\beta})$ in closed form as a function of $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{y}$. (2 points)

(c) Suppose the y_i's were observed with differing variances. To be specific, suppose that

$$p(y_i | \boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma_i^2} \right), \tag{4}$$

i.e., y_i has mean $\boldsymbol{x}_i^T \boldsymbol{\beta}$ and variance σ_i^2, where the σ_i's are fixed, known, constants). Show that finding the maximum likelihood estimate of β is equivalent to solving a weight linear regression problem. State clearly what the w_i's are in terms of the σ_i's. (4 points)

5. To perform variable selection, three classical approaches were introduced in class, including variable subset selection, forward stepwise selection and backward stepwise selection.

(a) To deepen your understanding of these approaches, please make a table to describe their key procedures as well as the pros and cons. (6 points)

(b) Suppose we perform these three approaches on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \cdots, p$ predictors. **Explain** your answers:

- Which of the three models with k predictors has the smallest training RSS? (1 points)
- Which of the three models with k predictors has the smallest test RSS? (1 points)

(Note that: Solutions with the correct answer but without adequate explanation will not earn credit.)

6. Refer to [1, Ex. 3.5]. Consider the ridge regression problem

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{5}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. Show that problem (5) is equivalent to the problem

$$\hat{\beta}^{\text{c}} = \underset{\beta^{\text{c}}}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0^{\text{c}} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \beta_j^{\text{c}})^2 + \lambda \sum_{j=1}^{p} (\beta_j^{\text{c}})^2 \right\}. \tag{6}$$

Give the correspondence between β^{c} and the original β in (5). Characterize the solution to this modified criterion. Moreover, show that a similar result holds for the least absolute shrinkage and selection operator (LASSO). (10 points)

7. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the LASSO may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or LASSO model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting. (2 points)

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$. (4 points)

(c) Write out the LASSO optimization problem in this setting. (2 points)

(d) Argue that in this setting, the LASSO coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions. (2 points)

8. Refer to [1, Ex. 3.30]. Consider the elastic-net optimization problem:

$$\min_{\beta} \ \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \left[\alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1 \right]. \tag{7}$$

Show how one can turn this into a LASSO problem, using an augmented version of $\boldsymbol{X}$ and $\boldsymbol{y}$. (10 points)

Optimization and Machine Learning, Spring 2020
Homework 2

(Due Wednesday, Apr. 1 at 11:59pm (CST))

March 27, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variable and one categorical response with K states.
   (a) Please define this multi-class classification problem, and solve it by ridge regression. (4 points)
   (b) Please make the prediction of a testing sample $x \in \mathbb{R}^p$ based on your model in (a). (3 points)
   (c) Is there any limitation on your model? If yes, please explain the problem by drawing a picture. (3 points)
   (d) Can you propose a model to overcome this limitation? If yes, please derive the decision boundary between an arbitrary class-pair. (5 points)
   (e) Can you revise your model in (d) by strength or weaken its assumptions? If yes, please tell the difference between your models in (d) and (e). (5 points)

2. Given an random variable, we have N i.i.d. observations by repeated experiments.
   (a) If the variable is boolean, please calculate the log-likelihood function. (4 points)
   (b) If the variable is categorical, please calculate the log-likelihood function. (4 points)
   (c) If the variable is continuous and follows Gaussian distribution, please calculate the log-likelihood function. (5 points)
   (d) Please discuss the difference between Maximum Likelihood Estimation (MLE) and Maximum a Posterior (MAP) estimation based on ONE of your results in (a), (b) and (c). (7 points)

3. Given the input variables $X \in \mathbb{R}^p$ and a response variable $Y \in \{0, 1\}$, the Expected Prediction Error (EPE) is defined by
$$\text{EPE} = \mathbb{E}[L(Y, \hat{Y}(X))],$$
where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{Y}(X))$ is a loss function measuring the difference between the estimated $\hat{Y}(X)$ and observed Y.
   (a) Given the zero-one loss
$$L(k, \ell) = \begin{cases} 1 & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell, \end{cases}$$
   please derive the Bayes classifier $\hat{Y}(x) = \text{argmax}_{k \in \{0,1\}} \Pr(Y = k | X = x)$ by minimizing EPE. (2 points)
   (b) Please define a function which enables to map the range of an arbitrary linear function to the range of a probability. (2 points)
   (c) Based on the function you defined in (b), please approximate the Bayes classifier in (a) by a linear function between X and Y, and derive its decision boundary. (4 points)
   (d) If each element of X is boolean, please show how many independent parameters are needed in order to estimate $\Pr(Y|X)$ directly; and is there any way to reduce its number? If yes, please describe your way mathematically. (4 points)
   (e) Based on your results in (d) and the Bayes theorem, please develop a classifier with a linear number of parameters w.r.t. p, and estimate these parameters by MLE. (5 points)
   (f) Please find at least three different points between your developed models in (c) and (e). (3 points)

4. Consider 12 labeled data points sampled from three distinct classes:

$$\text{Class 0}: \begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}-2\\0\end{bmatrix}, \begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}-3\\-5\end{bmatrix} \qquad \text{Class 1}: \begin{bmatrix}\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}-\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \begin{bmatrix}-4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \qquad \text{Class 2}: \begin{bmatrix}3\\5\end{bmatrix}, \begin{bmatrix}-1\\3\end{bmatrix}, \begin{bmatrix}8\\6\end{bmatrix}, \begin{bmatrix}0\\-2\end{bmatrix}$$

   (a) For each class $C \in [0, 1, 2]$, compute the class sample mean μ_C, the class sample covariance matrix Σ_C, and the estimate of the prior probability π_C that a point belongs to class C. (6 points)

   (b) Suppose that we apply LDA to classify the data given in part (a). Will this get the good decision boundary? Briefly explain your answer. (4 points)

5. We have two classes, named N for normal and E for exponential. For the former class ($Y = N$), the prior probability is $\pi_N = P(Y = N) = \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}$ and the class conditional $P(X|Y = N)$ has the normal distribution $N(0, \sigma^2)$. For the latter, the prior probability is $\pi_E = P(Y = E) = \frac{1}{1+\sqrt{2\pi}}$ and the class conditional has the exponential distribution.

$$P(X = x|Y = E) = \begin{cases}\lambda e^{-\lambda x} & \text{if } x \geq 0\\ 0 & \text{if } x < 0\end{cases}$$

   Write an equation in x for the decision boundary. (Only the positive solutions of your equation will be relevant; ignore all $x < 0$.) Simplify the equation until it is quadratic in x. (You don't need to solve the quadratic equation. It should contain the constants σ and λ. Ignore the fact that 0 might or might not also be a point in the decision boundary.) (10 points)

6. Given data $\{(x_i, y_i) \in R^d \times \{0, 1\}\}_{i=1}^n$ and a query point x, we choose a parameter vector θ to minimize the loss (which is simply the negative log likelihood, weighted appropriately):

$$l(\theta; x) = -\sum_{i=1}^{n} w_i(x)[y_i \log(\mu(x_i)) + (1 - y_i) \log(1 - \mu(x_i))]$$

   where

$$\mu(x_i) = \frac{1}{1 + e^{-\theta \cdot x_i}}, w_i(x) = \exp(-\frac{||x - x_i||^2}{2\tau})$$

   where τ s a hyperparameter that must be tuned. Note that whenever we receive a new query point x, we must solve the entire problem again with these new weights $w_i(x)$.

   (a) Given a data point x, derive the gradient of $l(\theta; x)$ with respect to θ. (4 points)

   (b) Given a data point x, derive the Hessian of $l(\theta; x)$ with respect to θ. (4 points)

   (c) Given a data point x, write the update formula for Newton's method. (2 points)

7. Now we discuss Bayesian inference in coin flipping. Let's denote the number of heads and the total number of trials by N_1 and N, respectively.

   (a) Please derive the MAP estimation based on the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. (4 points)

   (b) Please derive the MAP estimation based on the following prior:

$$p(\theta) = \begin{cases}0.5 & \text{if } \theta = 0.5\\ 0.5 & \text{if } \theta = 0.4\\ 0 & \text{otherwise,}\end{cases}$$

   that believes the coin is fair, or is slightly biased towards tails. (4 points)

   (c) Suppose the true parameter is $\theta = 0.41$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large? (2 points)

Optimization and Machine Learning, Spring 2020
Homework 3
(Due Tuesday, Apr. 28 at 11:59pm (CST))

1. (a) Consider the linear regression from a probabilistic perspective. Suppose we are given a set of N observations of the input vector $\mathbf{x}$, which we denote collectively by a data matrix $\mathbf{X}$ whose n-th row is $\mathbf{x}_n^T$ with $n = 1, \cdots, N$. The corresponding target values are $\boldsymbol{t} = (t_1, \cdots, t_N)^T$. We can express uncertainty over the value of target variable using a probability distribution. Assume that given the data $\mathbf{x}_n$ and coefficient vector $\mathbf{w}$, the corresponding value of t_n has a Gaussian distribution with variance σ^2. If the data are assumed to be drawn independently, then the likelihood function is given by

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2). \tag{1}$$

Next we similarly introduce a prior distribution over the parameter vector $\mathbf{w}$, we shall consider a zero-mean Gaussian prior with variance α_i for each w_i. Assume that the parameter variables are independent. Thus the parameter prior takes the form

$$p(\mathbf{w} \mid \alpha) = \prod_{n=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1}). \tag{2}$$

Draw a directed probabilistic graphical model corresponding to the relevance vector machine described by equations (1) and (2). (5 points)

(b) Consider the model defined in (a). Suppose we are given a new input data $\hat{x}$ and we wish to find the corresponding probability distribution for $\hat{t}$ conditioned on the observed data. The graphical model that describes this problem is shown in following Fig. 1. Please give the corresponding joint distribution of all of the random variables in this model and conditioned on the deterministic parameters, i.e., $p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2)$. (5 points)

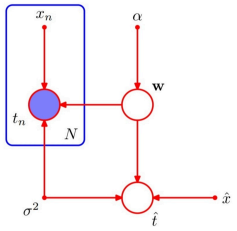


Figure 1: The graphical model.

2. According to the following Fig. 2, use the D-separation to analyze the following cases:

(a) Given x_4, $\{x_1, x_2\}$ and $\{x_6, x_7\}$ are conditionally independent. (5 points)

(b) Given $\{x_6, x_7\}$, x_3 and x_5 are conditionally independent. (5 points)

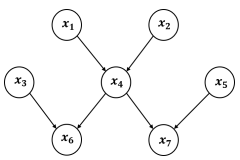


Figure 2: The Bayesian network for questions 2 and 3.

3. According to the Fig. 2, if all the nodes are observed and boolean variables, please complete the process of learning the parameter $\theta_{x_4|i,j}$ by using **MLE**, where $\theta_{x_4|i,j} = p(x_4 = 1 \mid x_1 = i, x_2 = j), i, j \in \{0, 1\}$. (15 points)

4. Define a Bayesian network with five discrete variables, represented by $\{F, A, S, H, N\}$. $\{F, A, H, N\}$ are 0/1 binary variables and $S \in \{0, 1, 2\}$, as illustrated in Fig. 3. Among them, $\{F, A, N\}$ are observed variables and $\{S, H\}$ are latent variables. Now we implement EM algorithm for this model.

   (a) If all five variables are observed, derive MLE of this model. You should state the close-form solution for each parameter you define. (5 points)

   (b) At least how many parameters should be defined for EM algorithm? (2 points)

   (c) Derive the E-step. You should enumerate each term. (4 points)

   (d) Derive the M-step. (4 points)

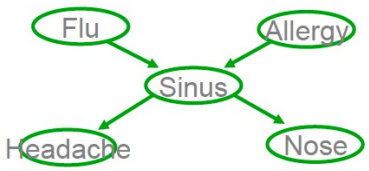


Figure 3: The Bayesian network for question 4.

5. Consider a set of K binary variables x_i, where $i = \{1, ..., K\}$, each variable $x_i \sim Bern(\mu_i)$. So $P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i}(1 - \mu_i)^{1-x_i}$, where $\mathbf{x} = (x_1, ..., x_K)^T$ *and* $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)^T$. The mean and covariance of this distribution are easily seen to be $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\text{cov}[\mathbf{x}] = \text{diag}\{\mu_i(1 - \mu_i)\}$.

   Now define a finite mixture of N Bernoullis given by $P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \pi_n P(\mathbf{x}|\boldsymbol{\mu_n})$ where $\boldsymbol{\mu} = \{\boldsymbol{\mu_1}, ..., \boldsymbol{\mu_N}\}$, $\boldsymbol{\pi} = \{\pi_i, ..., \pi_N\}$ and $P(\mathbf{x}|\boldsymbol{\mu_n}) = \prod_{i=1}^{K} \mu_{ni}^{x_i}(1 - \mu_{ni})^{1-x_i}$.

   (a) Derive the mean of the mixture distribution. (5 points)

   (b) Show the covariance of the mixture distribution equals $\sum_{n=1}^{N} \pi_n\{\boldsymbol{\Sigma_n} + \boldsymbol{\mu_n}\boldsymbol{\mu_n}^T\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T$, where $\boldsymbol{\Sigma_n} = \text{diag}\{\mu_{ni}(1 - \mu_{ni})$. (5 points)

6. Derive EM algorithm for the mixture of Bernoulli distributions above. There are D data points in total, where $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_D}\}$. (15 points)

7. Hoeffding's inequality is a powerful technique—perhaps the most important inequality in learning theory for bounding the probability that sums of bounded random variables are too large or too small. Below are some related inequalities you are required to provide proof:

   (a) **(Markov's inequality).** Let $Z \geq 0$ be a non-negative random variable. Then for all $t \geq 0$, show that

   $$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}(Z)}{t}, \tag{3}$$

   where $\mathbb{E}$ denotes the expectation operator. (6 points)

   (b) **(Chebyshev's inequality).** Let $Z \geq 0$ be a random variable with $\text{Var}(Z) < \infty$. Show that

   $$\mathbb{P}(Z \geq \mathbb{E}[Z] + t \text{ or } Z \leq \mathbb{E}[Z] - t) \leq \frac{\text{Var}(Z)}{t^2}, \qquad \text{for } t \geq 0, \tag{4}$$

   where $\text{Var}(Z)$ denotes the variance of Z. (6 points)

8. Recall that to show VC dimension is d for hypotheses $\mathcal{H}$ can be done via showing that VC dim$(\mathcal{H}) \leq d$ and VC dim$(\mathcal{H}) \geq d$. More specifically, to prove that VC dim$(\mathcal{H}) \geq d$ it suffices to give d examples that can be shattered; to prove VC dim$(\mathcal{H}) \leq d$ one must show that no set $d + 1$ examples can be shattered.

   For each one of the following function classes, find the VC dimension. State your reasoning based on the presented hint above. (Note that: solutions with the correct answer but without adequate explanation will not earn marks.)

(a) **Halfspaces in $\mathbb{R}^2$.** Examples lying in or on the halfspace are labeled $+1$, and the remaining examples are labeled -1. (3 points)

(b) **Axis-parallel rectangles in $\mathbb{R}^2$.** Points lying on or inside the target rectangle are labeled $+1$, and points lying outside the target rectangle are labeled -1. (3 points)

(c) **Closed sets in $\mathbb{R}^2$.** All points lying in the set or on the boundary of the set are labeled $+1$, and all points lying outside the set are labeled -1. (3 points)

(d) How many training examples suffice to assure with probability 0.9 that a consistent learner using the function classes presented in (b) will learn the target function with accuracy of at least 0.95? (4 points) (Hint: we use the following bounds on sample complexity: $m \geq \frac{1}{\epsilon}(4\log_2(2/\delta)+8\text{VC dim}(\mathcal{H})\log_2(13/\epsilon))$).

Optimization and Machine Learning, Spring 2020
Homework 4
(Due Tuesday, May 12 at 11:59pm (CST))

1. Given a training dataset $S = \{(x_i, y_i)\}_{i=1}^n$, in which $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ denote the i-th sample and the i-th label, respectively. Suppose that we use S to train a machine learning model based on Adaboost. At the end of the t-th iteration ($t = 1, 2, ..., T$), the importance of the i-th ($i = 1, 2, ..., n$) sample x_i is reweighted as

$$D_i^{(t+1)} = D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right),$$

where α_t is the weight of the t-th weakly binary classifier h_t, i.e.,

$$\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t}), \text{ with } \epsilon_t = \sum_{i=1}^{n} D_i^{(t)} \mathbb{1}(y_i \neq h_t(x_i)).$$

To classify an arbitrary test sample x, we calculate $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ and then return its sign. Now let's show that if every learner h_t ($\forall t$) achieves 51% classification accuracy (that is, only slightly better than random guessing), AdaBoost will converge to zero training error.

   (a) Let's change the update rule so that the weights of each iteration are normalized, that is, $\sum_{i=1}^{n} D_i^{(t)} = 1$ ($\forall t$). In this sense, we can treat the weights as a discrete probability distribution over the sample points. Hence we rewrite the update rule by

$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

   where Z_t is the normalization factor, $\forall t$. Please show that the following formula is satisfied,

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}.$$

   (5 points)

   (b) Assume that the initial weights follow uniform distribution, i.e.,

$$D_1^{(1)} = D_2^{(1)} = \cdots = D_n^{(1)} = \frac{1}{n}.$$

   Please show that

$$D_i^{(T+1)} = \frac{1}{n\prod_{t=1}^{T} Z_t} e^{-y_i f(x_i)},$$

   where $i = 1, 2, ..., n$ and $t = 2, 3, ..., T$. (5 points)

   (c) Let m be the number of sample points that Adaboost classifies incorrectly. Please show that

$$\sum_{i=1}^{n} e^{-y_i f(x_i)} \geq m.$$

   (5 points)

   (d) Based on the results in (a), (b), and (c), please show that once $\epsilon_t \leq 0.49$ is satisfied for every learner h_t ($\forall t$), then we have $m \to 0$ as $T \to \infty$. (5 points)

2. Suppose that we are interested in learning a classifier, such that at any turn of a game we can pose a question, like "should I attack this ant hill now?", and get an answer. That is, we want to build a classifier which we can feed some features on the current game state, and get the output "attack" or "don't attack". There are many possible ways to define what the action "attack" means, but for now let's define it as sending all friendly ants that can see the ant hill under consideration towards it.

Let's recall the AdaBoost algorithm described in class. Its input is a dataset $\{(x_i, y_i)\}_{i=1}^n$, with x_i being the i-th sample, and $y_i \in \{-1, 1\}$ denoting the i-th label, $i = 1, 2, ..., n$. The features might be composed of a count of the number of friendly ants that can see the ant hill under consideration, and a count of the number of enemy ants these friendly ants can see. For example, if there were 10 friendly ants that could see a particular ant hill, and 5 enemy ants that the friendly ants could see, we would have:

$$x_1 = \begin{bmatrix} 10 \\ 5 \end{bmatrix}.$$

The label of the example x_1 is $y_1 = 1$, once the friendly ants were successful in razing the enemy ant hill, and $y_1 = 0$ otherwise. We could generate such examples by running a greedy bot (or any other opponent bot) against a bot that we make periodically try to attack an enemy ant hill. Each time this bot tries the attack, we record (say, after 20 turns or some other significant amount of time) whether the attack was successful or not.

(a) Let ϵ_t denote the error of a weak classifier h_t:

$$\epsilon_t = \sum_{i=1}^{n} D_i^{(t)} \mathbb{1}(y_i \neq h_t(x_i)).$$

In the simple "attack" / "don't attack" scenario, suppose that we have implemented the following six weak classifiers:

$$h^{(1)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 2) - 1, \qquad h^{(4)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 2) - 1,$$
$$h^{(2)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 5) - 1, \qquad h^{(5)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 5) - 1,$$
$$h^{(3)}(x_i) = 2 * \mathbb{1}(x_{i1} \geq 10) - 1, \qquad h^{(6)}(x_i) = 2 * \mathbb{1}(x_{i2} \leq 10) - 1.$$

Given ten training data points ($n = 10$) as shown in Fig. 1, please show that what is the minimum value

i	x_{i1}	x_{i2}	y_i
1	3.5	0.5	1
2	2.5	1.5	1
3	5.5	5.5	1
4	7.5	7.5	1
5	6.5	9.5	1
6	1.5	2.5	-1
7	4.5	4.5	-1
8	3.5	7.5	-1
9	1.5	11.5	-1
10	6.5	11.5	-1

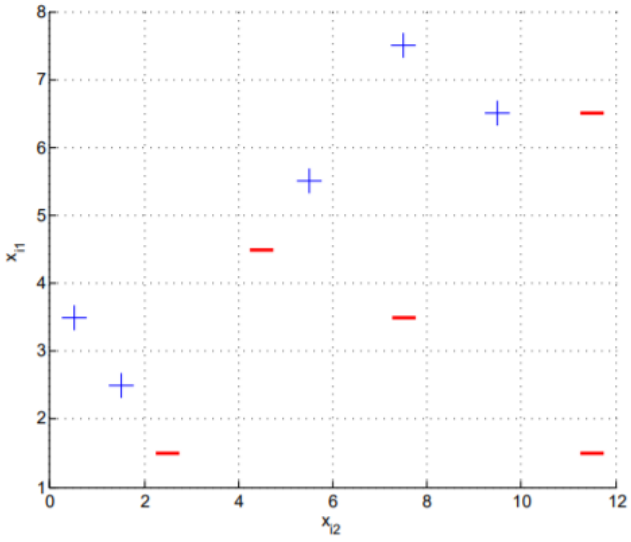


Figure 1: The training data in (a).

of ϵ_1 and which of $h^{(1)}, ..., h^{(6)}$ achieve this value? Note that there may be multiple classifiers that all have the same ϵ_1. You should list all classifiers that achieve the minimum ϵ_1 value. (5 points)

(b) For all the questions in the remainder of this section, let h_1 denote $h^{(1)}$ chosen in the first round of boosting. (That is, $h^{(1)}$ was the classifier that achieved the minimum ϵ_1.)

(1) What is the value of α_1 (the weight of this first classifier h_1)? Keep in mind that the log in the formula for α_t is a natural log (base e). (5 points)

(2) What should Z_t be in order to make sure the distribution $D^{(t+1)}$ is normalized correctly? That is, derive the formula of Z_t in terms of $D^{(t)}$, α_t, h_t, and $\{(x_i, y_i)\}_{i=1}^n$, that will ensure $\sum_{i=1}^n D_i^{(t+1)} = 1$. (5 points)

(3) Which points will increase in significance in the second round of boosting? That is, for which points will we have $D_i^{(1)} < D_i^{(2)}$? What are the values of $D^{(2)}$ for these points? (5 points)

(4) In the second round of boosting, the weights on the points will be different, and thus the error ϵ_2 will also be different. Which of $h^{(1)}, ..., h^{(6)}$ will minimize ϵ_2? (Which classifier will be selected as the second weak classifier h_2?) What is its value of ϵ_2? (5 points)

(5) What will the average error of the final classifier H be, if we stop after these two rounds of boosting? That is, if $H(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x))$, what will the training error $\epsilon = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i))$ be? Is this more, less, or the same as the error we would get, if we just used one of the weak classifiers instead of this final classifier H? (5 points)

3. Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, please verify the following new kernels will also be valid:

   (a) $k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}')$, where $f(\cdot)$ is any function. (2 points)

   (b) $k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}'))$, where $q(\cdot)$ is a polynomial with nonnegative coefficients. (3 points)

   (c) $k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}'))$. (5 points)

   (d) $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}'$, where $\mathbf{A}$ is a symmetric positive semi-definite matrix. (5 points)

4. Consider the space of all possible subsets A of a given fixed set D. Show that the kernel function $k(A_1, A_2) = 2^{|A_1 \cap A_2|}$ corresponds to an inner product in a feature space of dimensionality $2^{|D|}$ defined by the mapping $\phi(A)$ where A is a subset of D and the element $\phi_U(A)$, indexed by the subset U, is given by

$$\phi_U(A) = \begin{cases} 1, & \text{if } U \subseteq A; \\ 0, & \text{otherwise.} \end{cases}$$

   Here $U \subseteq A$ denotes that U is either a subset of A of is equal to A. (10 points)

5. Suppose we have a data set of input vectors $\{\mathbf{x}_n\}$ with corresponding target values $t_n \in \{-1, 1\}$, and suppose that we model the density of input vectors within each class separately using a Parzen kernel density estimator which is defined as follows

$$p(\mathbf{x}|t) = \frac{1}{N_t} \sum_{n=1}^{N} \frac{1}{Z_k} k(\mathbf{x}, \mathbf{x}_n) \delta(t, t_n)$$

   where $k(\mathbf{x}, \mathbf{x}')$ is a valid kernel, Z_k is the normalization constant for the kernel and $\delta(t, t_n)$ equals 1 if $t = t_n$ and 0 otherwise.

   (a) Write down the minimum misclassification-rate decision rule assuming the two classes have equal prior probability. (3 points)

   (b) Show that, if the kernel is chosen to be $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$, then the classification rule reduces to simply assigning a new input vector to the class having the closest mean. (4 points)

   (c) Show that, if the kernel takes the form $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$, that the classification is based on the closest mean in the feature space $\phi(\mathbf{x})$. (4 points)

6. The problem of maximizing margin can be converted into an following equivalent problem

$$\begin{aligned} &\underset{\mathbf{w}, b}{\text{argmin}} \ \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to } t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1, \quad n = 1, \cdots, N. \end{aligned}$$

   where $\phi(\mathbf{x})$ is a fixed feature-space transformation.

   (a) By introducing Lagrange multipliers $\{a_n\}$, please give the Lagrangian function and the dual representation of the maximum margin problem. (8 points)

   (b) Please show that the value ρ of the margin for the maximum-margin hyperplane is given by

$$\frac{1}{\rho^2} = \sum_{n=1}^{N} a_n.$$

   (Hint: $\{a_n\}$ can be obtained by solving the dual representation of the maximum margin problem.) (6 points)

Optimization and Machine Learning, Spring 2020
Homework 5
(Due Tuesday, June 2 at 11:59pm (CST))

1. Show that weighted Euclidean distance in $\mathbb{R}^p$,

$$d_e^{(w)}(x_i, x_{i'}) = \frac{\sum_{l=1}^p w_l (x_{il} - x_{i'l})^2}{\sum_{l=1}^p w_l},$$

satisfies

$$d_e^{(w)}(x_i, x_{i'}) = d_e(z_i, z_{i'}) = \sum_{l=1}^p (z_{il} - z_{i'l})^2,$$

where

$$z_{il} = x_{il} \left(\frac{w_l}{\sum_{l=1}^p w_l} \right)^{1/2}.$$

Thus weighted Euclidean distance based on x is equivalent to unweighted Euclidean distance based on z. (15 points)

2. Consider a dataset of n observations $\mathbf{X} \in \mathbb{R}^{n \times d}$, and our goal is to project the data onto a subspace having dimensionality p, $p < d$. Prove that PCA based on projected variance maximization is equivalent to PCA based on projected error (Euclidean error) minimization. (20 points)

3. Show that the conventional linear PCA algorithm is recovered as a special case of kernel PCA if we choose the linear kernel function given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$. (15 points)

4. Let $S = \{x^{(1)}, \cdots, x^{(n)}\}$ be a dataset of n samples with 2 features, i.e $x^{(i)} \in \mathbb{R}^2$. The samples are classified into 2 categories with labels $y^{(i)} \in \{0, 1\}$. A scatter plot of the dataset is shown in Figure 1.

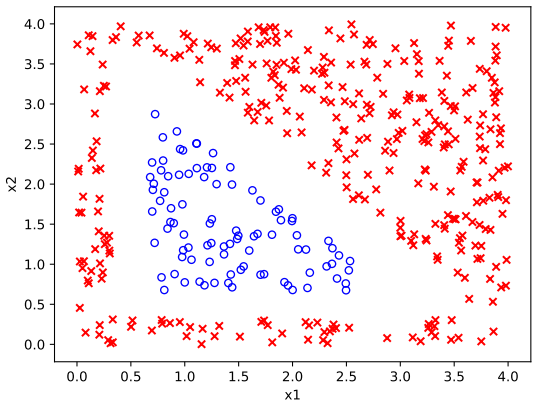


Figure 1: Plot of dataset S.

The examples in class 1 are marked as as "$\times$" and examples in class 0 are marked as "$\circ$". We want to perform binary classification using a simple neural network with the architecture shown in Figure 2.

Denote the two features x_1 and x_2, the three neurons in the hidden layer h_1, h_2, and h_3, and the output neuron as o. Let the weight from x_i to h_j be $w_{i,j}^{[1]}$ for $i \in \{1, 2\}, j \in \{1, 2, 3\}$, and the weight from h_j to o be $w_j^{[2]}$. Finally, denote the intercept weight for h_j as $w_{0,j}^{[1]}$, and the intercept weight for o as $w_0^{[2]}$. For the loss

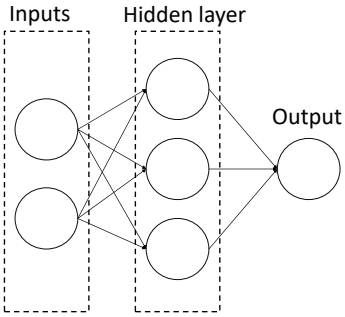


Figure 2: Architecture for our simple neural network.

function, we'll use average squared loss instead of the usual negative log-likelihood:

$$l = \frac{1}{n}\sum_{i=1}^{n}(o^{(i)} - y^{(i)})^2,$$

where $o^{(i)}$ is the result of the output neuron for example i.

(a) Suppose we use the sigmoid function as the activation function for h_1, h_2, h_3 and o. What is the gradient descent update to $w_{2,1}^{[1]}$, assuming we use a learning rate of α? Your answer should be written in terms of $x^{(i)}$, $o^{(i)}$, $y^{(i)}$, and the weights. (10 points)

(b) Now, suppose instead of using the sigmoid function for the activation function for h_1, h_2, h_3 and o, we instead used the step function $f(x)$, defined as

$$f(x) = \begin{cases} 1, x \geq 0, \\ 0, x < 0. \end{cases}$$

Is it possible to have a set of weights that allow the neural network to classify this dataset with 100% accuracy? If it is possible, please provide a set of weights that enable 100% accuracy and explain your reasoning for those weights in your PDF. If it is not possible, please explain your reasoning in your PDF. (10 points) (Hint: There are three sides to a triangle, and there are three neurons in the hidden layer.)

(c) Let the activation functions for h_1, h_2, h_3 be the linear function $f(x) = x$ and the activation function for o be the same step function as before. Is it possible to have a set of weights that allow the neural network to classify this dataset with 100% accuracy? If it is possible, please provide a set of weights that enable 100% accuracy and explain your reasoning for those weights in your PDF. If it is not possible, please explain your reasoning in your PDF. (10 points)

5. Convolutional neural networks targets the processing of 2-D features instead of the 1-D ones in multi-layer perceptron (MLP), the structure of which is depicted in Fig. 3.

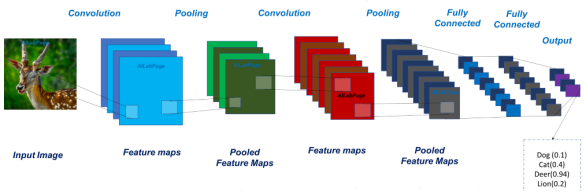


Figure 3: https://mc.ai/how-does-convolutional-neural-network-work/

(a) Kernel convolution is process where we take a kernel (or filter), we pass it over our image and transform it based on the values from filter. Now, you are given the following formula

$$G(m,n) = (f * h)(m,n) = \sum_j \sum_k h(j,k) f(m-j, n-k),$$

where the input image is denoted by f and the kernel by h. The indexes of rows and columns of the result matrix are marked with m and n respectively. Please calculate the feature maps, if you are given the following 3×3 image matrix and 2×2 kernel matrix. (5 points)

1	2	3
4	5	6
7	8	9

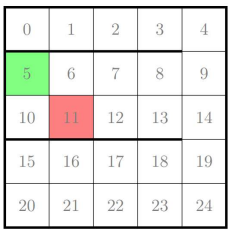
Table 1: 3×3-Image Matrix.

1	0
0	1

Table 2: 2×2-Kernel Matrix.

(b) Assume the input with the size of (width = 28, height = 28) and a filter with the size of (width =5, height = 5) and the convolutional layer parameters are $S = 1$ (the stride), $P = 0$ (the amount of zero padding). What is the exact size of the convolution output? (5 points)

6. Consider the following grid environment. Starting from any unshaded square, you can move up, down, left, or right. Actions are deterministic and always succeed (e.g. going left from state 1 goes to state 0) unless they will cause the agent to run into a wall. The thicker edges indicate walls, and attempting to move in the direction of a wall results in staying in the same square. Taking any action from the green target square (no. 5) earns a reward of +5 and ends the episode. Taking any action from the red square of death (no. 11) earns a reward of -5 and ends the episode. Otherwise, each move is associated with some reward $r \in \{-1, 0, +1\}$. Assume the discount factor $\gamma = 1$ unless otherwise specified.



(a) Define the reward r for all states (except state 5 and state 11 whose rewards are specified above) that would cause the optimal policy to return the shortest path to the green target square (no. 5). (3 points)

(b) Using r from part (a), find the optimal value function for each square. (5 points)

(c) Does setting $\gamma = 0.8$ change the optimal policy? Why or why not? (2 points)

Optimization and Machine Learning, Spring 2020
Homework 6
(Due Tuesday, June 16 at 11:59pm (CST))

1. Which of the following sets are convex?
   (a) A *wedge, i.e.*, $\{x \in \mathbb{R}^n \mid a_1^T x \leq b_1, a_2^T x \leq b_2\}$. (5 points)
   (b) The set of points closer to a given point than a given set, *i.e.*,
   $$\{x \mid ||x - x_0||_2 \leq ||x - y||_2 \text{ for all } y \in S\}$$
   where $S \subseteq \mathbb{R}^n$. (5 points)
   (c) The set of points closer to one set than another, *i.e.*,
   $$\{x \mid \mathbf{dist}(x, S) \leq \mathbf{dist}(x, T)\}$$
   where $S, T \subseteq \mathbb{R}^n$, and
   $$\mathbf{dist}(x, S) = \inf\{||x - z||_2 \mid z \in S\}.$$
   (5 points)
   (d) The set $\{x \mid x + S_2 \subseteq S_1\}$, where $S_1, S_2 \subseteq \mathbb{R}^n$ with S_1 convex. (5 points)
   (e) The set of multiplication
   $$\{x \in \mathbb{R}^n_+ \mid \prod_{i=1}^n x_i \geq 1\}.$$
   (5 points)

2. Determine whether the following functions are convex, strictly convex, concave, strictly concave, both or neither.
   (a) $f(x) = e^x - 1$ on $\mathbb{R}$. (5 points)
   (b) $f(x_1, x_2) = x_1 x_2$ on $\mathbb{R}^2_{++}$. (5 points)
   (c) $f(x) = \log(\sum_{i=1}^n \exp(x_i))$ on $\mathbb{R}^n$, use the second-order condition. (5 points)
   (d) $f(w) = ||Xw - y||_2^2 + \lambda||w||_2^2$ for $\lambda > 0$. (5 points)
   (e) The log-likelihood of a set of points $\{x_1, \cdots, x_n\}$ that are normally distributed with mean μ and finite variance $\sigma > 0$ is given by:
   $$f(\mu, \sigma) = n \log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$
   Show that if we view the log likelihood for fixed σ as a function of the mean, *i.e.*,
   $$g(\mu) = n \log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$
   then g is strictly concave.
   Show that if we view the log likelihood for fixed μ as a function of z, *i.e.*,
   $$h(z) = n \log(\frac{\sqrt{z}}{\sqrt{2\pi}}) - \frac{z}{2} \sum_{i=1}^n (x_i - \mu)^2$$
   then h is strictly concave (equivalently, we say f is strictly concave in $z = \frac{1}{\sigma^2}$).
   We say $f(x, y)$ with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ is jointly convex if
   $$f\big(\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2)\big) \leq \lambda f\big((x_1, y_1)\big) + (1 - \lambda) f\big((x_2, y_2)\big).$$
   Show that f is not jointly concave in $\mu, \frac{1}{\sigma^2}$. (5 points)

3. Consider the problem

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|_1/\left(c^Tx+d\right) \\ \text{subject to} & \|x\|_\infty \leq 1, \end{array}$$

where $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$ and $d \in \mathbb{R}$. We assume that $d > \|c\|_1$, which implies that $c^{\mathrm{T}}x + d > 0$ for all feasible x.

(a) Show that this is a quasiconvex optimization problem. (5 points)

(b) Show that it is equivalent to the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \|Ay-bt\|_1 \\ \text{subject to} & \|y\|_\infty \leq t, \\ & c^Ty+dt=1, \end{array}$$

with variables $y \in \mathbb{R}^n, t \in \mathbb{R}$. (10 points)

4. Consider the QCQP

$$\begin{array}{ll} \text{minimize} & (1/2)x^TPx+q^Tx+r \\ \text{subject to} & x^Tx \leq 1, \end{array}$$

with $P \in \mathbf{S}^n_{++}$. Show that $x^\star = -(P+\lambda I)^{-1}q$ where $\lambda = \max\{0, \bar{\lambda}\}$ and $\bar{\lambda}$ is the largest solution of the nonlinear equation

$$q^{\mathrm{T}}(P+\lambda I)^{-2}q = 1.$$

(15 points)

5. Consider the inequality form LP

$$\begin{array}{ll} \text{minimize} & c^Tx \\ \text{subject to} & Ax \preceq b, \end{array}$$

with $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Let $w \in \mathbb{R}^m_+$. If x is feasible for the LP, *i.e.*, satisfies $Ax \preceq b$, then it also satisfies the inequality

$$w^TAx \leq w^Tb.$$

Geometrically, for any $w \succeq 0$, the halfspace $H_w = \{x \mid w^TAx \leq w^Tb\}$ contains the feasible set for the LP. Therefore if we minimize the objective c^Tx over the halfspace Hw we get a lower bound on p^*.

(a) Derive an expression for the minimum value of c^Tx over the halfspace H_w (which will depend on the choice of $w \succeq 0$). (5 points)

(b) Formulate the problem of finding the best such bound, by maximizing the lower bound over $w \succeq 0$. (5 points)

(c) Relate the results of (a) and (b) to the Lagrange dual of the LP. (10 points)