MATH 462 COMBINED HW FOR EXAM STUDY
VERSION November 23, 2023

ADAM M. OBERMAN

1.1. **Decision Trees: Examples and Theory, basic coding.**

Exercise 1.1. Apply the decision tree algorithm, using the information gain as the attribute test to learn a decision tree for the data which follows. Be sure to present the information gain corresponding each attribute.

The data set is comprised of three attributes (binary input features) A_1, A_2, and A_3 and one binary output, along with five labelled samples.

Sample	A_1	A_2	A_3	Output y
x_1	1	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	1	1	1	1
x_5	1	1	0	1

Exercise 1.2. For each of the following logical functions of the attributes A, B, C, draw the corresponding decision tree.

(1) Tree for $h(A, B, C) = A \vee (B \wedge C)$.
(2) Tree for $h(A, B) = A\ XOR\ B$
(3) Tree for $h(A, B, C, D) = (A \wedge B) \vee (C \wedge D)$

Exercise 1.3 (Loss design). *For the definitions, refer to the notes, section 5, "Loss minimization for choosing attributes".*

(1) Define $\ell(p, 0) = cp^2$. Find the value of c so that the loss is normalized. Define $\ell(p, 1)$ so that the loss is balanced. Show that the loss rewards confidence.
(2) Suppose a dataset S has 10 positive and 4 negative examples. Evaluate the expression $L(q, S)$ for this loss. Find the value of the loss when $q = 10/14$, and when $q = .9$.

ANSWER: $\ell(p, 0) = 4p^2$, $\ell(p, 1) = 4(1 - p)^2$. derivative is nonzero.

Answer : $\frac{10}{14}4(1 - q)^2 + \frac{4}{14}4(q)^2$ When $q = 10/14$: simplifies to $4 * 10 * 4/14^2$

2.2. **Bayes.**

Exercise 2.4 (Bayes Rule Medical Test). *Refer to the page 3, section 3.2. of Lecture 3 Bayes Lecture notes, Example of Bayes rule.* Suppose we consider a medical test for a disease. We have prior knowledge that over the entire population of people only .0002 have this disease. The test returns a correct positive result in only 97% of the cases in which the disease is actually present and a correct negative result in only 95% of the cases in which the disease is not present. In other cases, the test returns the opposite result.

(1) Suppose a person, randomly selected from the population is tested, and the test comes back positive. Use Bayes rule to calculate the probability that they have the disease.
(2) Suppose a person comes from a higher risk group, where the probability of having the disease is 20 times higher. Suppose this person receives a positive rest result. What is the probability they have the disease?

Exercise 2.5 (Naive Bayes Classifier). (Refer to section 3 in Lecture 4 Bayes notes) The naive Bayes classifier defines

$$s(x) = w_0 + \sum_{i=1}^{K} w_i a_i(x)$$

where $a_i(x)$ is 1 when word i appears in x, and 0 otherwise, and where

$$w_i = \log\left(\frac{P(1 \mid a_i = 1)}{P(0 \mid a_i = 1)}\right).$$

Suppose that, in a given database, $P_1 = .65$ of the emails are spam, and the rest are ham (not spam), the word 'password' appears in 856 of the spam emails, and 230 of the ham emails, the word 'benefits' appears in 112 spam and 670 ham.

(1) Suppose email x contains the word 'password'. What is $s(x)$?
(2) Suppose email x contains both 'password' and 'benefits', what is $s(x)$?

2.3. **Neural Networks.**

Exercise 2.6. Consider a single layer perceptron, with the $\operatorname{sgn}(x)$ nonlinearity $h(x,w) = \operatorname{sgn}(w \cdot x)$. For binary data $x = (1, x_1, \dots x_n)$, solve the following problems. Find the weight vector $w = w_0, \dots, w_n$ for which $h(x,w)$ fits $y(x)$ (as best as possible). If possible, use 'simple' weight vectors, in other words choose w to make reading the function easier for a person.

(1) For $n = 2$, write out w for the $AND(x)$ function and for $OR(x)$.
(2) Write out w for the $AND(x)$ function and for $OR(x)$. for arbitrary $n \geq 2$.
(3) For $n > 2$, define $y(x) = 1$ if half or more of x_i are 1, and $y(x) = 0$ otherwise. Write out a w that fits this function for a perceptron.
(4) **coding** Implement these functions (with $n = 4$) in a perceptron code, and plot the decision boundary, as a function of x_1, x_2, when $x_3 = 0, x_4 = 1$.

2.4. **Vector Calculus for ML.**

Exercise 2.7. Let $f : \mathbb{R}^d \to \mathbb{R}$ and write $\|x\| = \sqrt{x_1^2 + \cdots + x_d^2}$. Find $\nabla f(x)$ for each of the following functions. Write the answer in vector notation.

(1) $f(x) = \|x\|^2$.
(2) $f(x) = \|x\|$,
(3) $f(x) = \|x - a\|$
(4) $f(x) = 1/\|x\|$
(5) $f(x) = \|m \cdot x - b\|^2$, where $m \in \mathbb{R}^d, b \in \mathbb{R}$
(6) $f(x) = \|m_1 \cdot x - b_1\|^2 + \dots \|m_n \cdot x - b_n\|^2$, where $m_i \in \mathbb{R}^d, b_i \in \mathbb{R}$

Exercise 2.8 (Vector calculus for loss minimization). Consider the following

(1) Let $w, x_1, \dots, x_m \in \mathbb{R}$. Define

$$L(w) = f(w, x_1, \dots, x_m) = \frac{1}{m}\sum_{i=1}^m (w - x_i)^2$$

Find $L'(w)$. Solve for $L'(w^*) = 0$. Evaluate $L(w^*)$ in terms of x.

(2) Now same problem, but with vector data. Let $w, x_1, \dots, x_m \in \mathbb{R}^d$. Set

$$L(w) = f(w, x_1, \dots, x_m) = \frac{1}{m}\sum_{i=1}^m \|w - x_i\|_2^2$$

Find $\nabla_w f(w, x_1, \dots, x_m)$. Write it in terms of $\bar{x} = \frac{1}{m}\sum_{i=1}^m x_i$ Find

$$w^* = \arg\min_w L(w), \qquad L(w^*) = \frac{1}{m}\sum_{i=1}^m \|w^* - x_i\|^2$$

(3) Interpret $w^*, L(w^*)$ in terms of familiar statistics (mean, variance).

3.5. **Losses.**

Exercise 3.9. Consider the normalized log loss,

$$\ell_{\log,2}(s,y) = \frac{1}{\log 2}\log(1+\exp(-ys))$$

(1) Plot it, (do one plot for each value of $y \in \pm 1$), and show in the plot that it is an upper bound for the zero-one loss.
(2) Show that it is balanced, normalized, monotone, and convex.

Exercise 3.10. Consider the exponential loss $\ell(s,y) = \exp(-ys)$.

(1) Plot it, (do one plot for each value of $y \in \pm 1$), and show in the plot that it is an upper bound for the zero-one loss.
(2) Show that it is balanced, normalized, monotone, and convex.

Exercise 3.11 (Multi-class losses). (1) Let $s = (5.2, 3, .5)$, calculate the $K = 3$ class $\log$ loss when $y = 1$ and when $y = 2$.
(2) Let $s = (5.2, 3, .5)$, calculate the $K = 3$ class margin loss when $y = 1$ and when $y = 2$.
(3) Let ℓ be (a) the multiclass log loss and (b) the multiclass margin loss. In each case, plot the function $\ell(s(t), y = 3)$ for $s(t) = (1, 2, t)$ with $t \in [-4, 4]$.

3.6. **Gradient Descent and SGD.**

Exercise 3.12 (Convergence rates and log plots). We say an algorithm converges exponentially with rate $c < 1$ if the error $e(t)$ satisfies $\log e(t)/e(0) \leq ct$. Consider the sequence $a(t) = 25(2/3)^t$.

(1) Show that it convergences exponentially and find the rate.
(2) Plot $a(t)$ a log-plot, so that the slope shows the rate of convergence. The x-axis should be the iteration count, and the y-axis should be the $\log$ of the error.
(3) Combine the previous plot with a log-plot for the sequences $(.99)^t$, $100(.99)^t$ and $.04^t$.

Exercise 3.13 (Gradient Descent and SGD Implementation). You may use the code provided as a starting point, or write your own.

`https://colab.research.google.com/drive/1-YoLDf3OyH3SxLJtC5W4qG3L1zYxkyMf?usp=sharing`

Consider the model problem, for $w \in \mathbb{R}$,

$$L(w) = \frac{1}{m}\sum_{i=1}^{m}\frac{(w-y_i)^2}{2}$$

for $w \in \mathbb{R}$, where $m = 500$ and y_i are uniformly generated over $[-1, 1]$.

(a) Run Gradient Descent on the model problem with learning rates $\alpha = .95, .75, .5, .25, .1$. In this case, you know the exact w^*. Plot the error, $e(w^t) = \|w^t - w^*\|^2$, on a log-plot, so that the slope shows the rate of convergence. The x-axis should be the iteration count, and the y-axis should be the $\log$ of the error, see sample below.
(b) Run the SGD algorithm, corresponding to example 2.3 in the notes. (Drawing balls with replacement). Consider a data set with $R = 10$ red ball and $B = 15$ blue balls, and let p_t be the estimate of the fraction of red balls. Do the update,

$$p_{t+1} = p_t - \frac{1}{t+1}(p_t - y_t)$$

where y_t is 1 if the ball is red, and zero otherwise. Plot the error, $e(t) = (p_t - p^*)^2$ as a function of t.

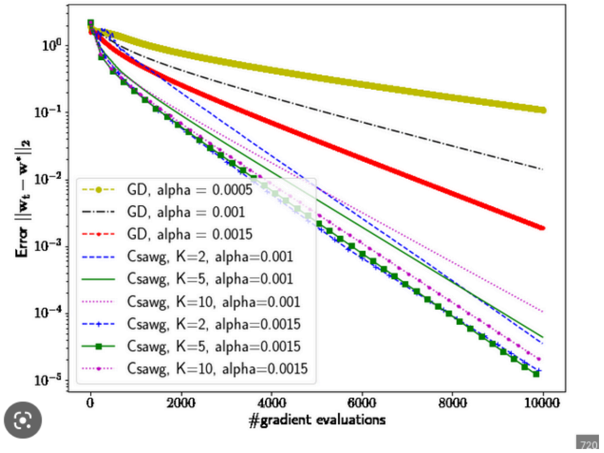

FIGURE 1. Sample convergence rate plot for gradient descen

4.7. **Analytic Geometry and Covariance matrices.**

Exercise 4.14 (Exercise 3.5 from MLL[1]). Consider the Euclidean vector space $\mathbb{R}^5$. A subspace $U \subseteq \mathbb{R}^5$ and $\boldsymbol{x} \in \mathbb{R}^5$ are given by

$$U = \operatorname{span}\left\{ \begin{bmatrix} 0 \\ -1 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \\ 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} -3 \\ 4 \\ 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -3 \\ 5 \\ 0 \\ 7 \end{bmatrix} \right\}, \quad \boldsymbol{x} = \begin{bmatrix} -1 \\ -9 \\ -1 \\ 4 \\ 1 \end{bmatrix}$$

(a) Determine the orthogonal projection $\pi_U(\boldsymbol{x})$ of $\boldsymbol{x}$ onto U
(b) Determine the distance $d(\boldsymbol{x}, U)$

Exercise 4.15 (Inner products). (a) Rewrite the definition of an inner product on a vector space.
(b) Given the $n \times n$ matrix M, which is full rank, verify from the definition that $\langle x, y \rangle_M = (Mx)^\top (My)$ defines an inner product on $\mathbb{R}^n$. What goes wrong if the matrix has a non-trivial null-space?
(c) Give an example of a norm on $\mathbb{R}^n$ which does not come from an inner product.

Exercise 4.16 (Covariance Matrix). Let $n = 2$. Find the covariance matrix for the following datasets, S^m.
(a) $S^m = \{(1,1), (-1,-1), (1,0), (-1,0), (-1,1), (1,-1), (0,1), (0,-1)\}$
(b) $S^m = \{(t,t), (-t,-t), (1,0), (-1,0), (-1,1), (1,-1), (0,1), (0,-1)\}$, for any $t \in \mathbb{R}$.

Exercise 4.17 (Covariance Matrix Theory). Prove that the covariance matrix, C, is symmetric and non-negative definite, meaning $x^\top C x \geq 0$ for all x. Assuming that the matrix is invertible, prove it is (strictly) positive definite.

4.8. **k-means Clustering.**

Exercise 4.18. For the following two dimensional data sets, plot the data by hand and indicate the clusterings by drawing a circle around the points in each cluster

(a) The dataset

$$\begin{aligned} S = \{&(.8,1), (1.2,1), (1,.8), (1,1.2), (-.8,-1), (-1.2,-1), (-1,-.8), (-1,-1.2), \dots \\ &(-.8,1), (-1.2,1), (-1,.8), (-1,1.2), (.8,-1), (1.2,-1), (1,-.8), (1,-1.2)\}. \end{aligned}$$

(b) $S = \{(3,1), (-3,-1), (-3,1), (3,-1)\}$. Indicate two possible clusterings.

[1] Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for machine learning. Cambridge University Press, 2020.

Exercise 4.19. Consider the dataset $S = \{-5, -4, -3, 6, 8, 10\} \subset \mathbb{R}$.

(a) Starting from $W^0 = (w_1, w_2) = (6, 9)$ perform the k-means algorithm with $k = 2$ until a fixed point is reached.
(b) Let h_W be the minimizer of the k-means loss. Plot $h_W(x)$ on the interval $[-5, 10]$.

4.9. **Hypothesis classes for unsupervised ML.**

Exercise 4.20. Given a dataset $S = \{x_1, \dots, x_m\}$ in $\mathbb{R}^d$, which is mean zero, $\bar{x} = \frac{1}{m}\sum x_i = 0$. Given a hypothesis class $\mathcal{H}$ of functions $h : \mathbb{R}^d \to \mathbb{R}^d$, consider the loss $L(h, S) = \frac{1}{m}\sum_{i=1}^m \|h(x_i) - x_i\|^2$. In this exercise we identify algorithms in terms of this loss and a specific hypothesis class.

(a) Let $\mathcal{H}_{lin,k} = \{h(x) = W^\top W x \mid W \in \mathbb{R}^{k\times d}\}$. Show that this class includes the projection matrices onto rank k subspaces. How do you describe the W in this case? (Hint: orthgonality). If we minimize the loss $L(h, S)$ over this hypothesis class, which familiar algorithm do we obtain?
(b) Let $\mathcal{H} = \{h(x) = \arg\min_{i=1}^k \|x - w_i\|^2 \mid w_i \in \mathbb{R}^d, i = 1, \dots, k\}$. Describe the hypothesis class in words. Are the functions in $\mathcal{H}$ differentiable? Identify the corresponding algorithm.

5.10. **Convex Learning Problems.** Refer to Ch 12 of Understanding Machine Learning (Shalev-Shwartz).

For the next two exercises, consider the classification problem with $x \in X = [-1, 1]$, $y \in \{\pm 1\}$. Let $m = 6$ be the size of S, which is given by

$$S = \{(-1, -1), (-0.8, -1), (-0.6, +1), (-0.3, -1), (0.3, +1), (0.6, -1), (0.8, +1), (+1, +1)\}$$

Let $h(x, w) = x - w$, and $c(h) = \operatorname{sgn}(h)$.

Exercise 5.21 (non-convexity of 0-1 loss). With S and h as above, consider the zero-one loss

$$L(w) = \frac{1}{m}\sum_{i=1}^m \ell_{0-1}(c(x_i - w), y_i)$$

(a) Plot (sketch by hand) the function $L(w)$, for $w \in [-1, 1]$.
(b) Identify two local minima (they can be intervals), and the global minimum of the loss function.

Exercise 5.22 (Convex classification loss). With S and h as above, consider the classification loss $\ell_{\log}(h, y) = \log(1 + \exp(-yh))$, along with the loss function

$$L(w) = \frac{1}{m}\sum_{i=1}^m \ell_{\log}(h(x_i, w), y_i)$$

(a) Plot (or sketch by hand) the function $L(w)$, for $w \in [-1, 1]$.
(b) For the function $g(w, (x, y)) = \ell_{\log}(h, y)$, show that, for any values of (x, y) in the domain $X \times \{\pm 1\}$, g is a convex function of w. Find the first and second derivatives of g in w.
(c) Explain why L is also convex as a function of w.

Exercise 5.23 (Convexity, Lipschitz, and Smoothness of logistic regression loss.). Shalev-Shwartz Problem 12.2.

Exercise 5.24 (Lipschitz continuity of the hinge loss). Shalev-Shwartz Problem 12.3.

5.11. **Feature Regression and orthogonal features.**

Exercise 5.25. Consider the following vectors in $\mathbb{R}^3$:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

(a) Make these vectors orthogonal using the Gram-Schmidt process, by performing the following steps. (i) Start with the first vector and set $\mathbf{u_1} = \mathbf{v_1}$. (ii) Next, compute the projection of $\mathbf{v}_2$ onto $\mathbf{u}_1$ and subtract it from $\mathbf{v}_2$:

$$\mathbf{u}_2 = \mathbf{v}_2 - \mathsf{proj}_{\mathbf{u}_1}(\mathbf{v}_2), \quad \text{where } \mathsf{proj}_{\mathbf{u}_1}(\mathbf{v}_2) = \frac{\mathbf{u}_1 \cdot \mathbf{v}_2}{\|\mathbf{u}_1\|^2} \cdot \mathbf{u}_1$$

(iii) Lastly, compute the projection of $\mathbf{v}_3$ onto the subspace spanned by $\mathbf{u}_1$ and $\mathbf{u}_2$, and subtract it from $\mathbf{v}_3$:

$$\mathbf{u}_3 = \mathbf{v}_3 - \mathsf{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \mathsf{proj}_{\mathbf{u}_2}(\mathbf{v}_3)$$

Find the orthogonalized vectors $\mathbf{u}_1, \mathbf{u}_2$, and $\mathbf{u}_3$.

(b) Next, given the vector $y = [6, 0 - 6]^\top$. Solve

$$\min_w \|Fw - y\|^2, \quad F = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3], \quad Fw = \sum_i w_i \mathbf{u}_i$$

with F the 3×3 matrix with columns given by the vectors $\mathbf{u}_i$, . Express w_i in terms of inner products of y with certain vectors.

Exercise 5.26. Now consider the vector space of functions $u : X = [0,1] \to \mathbb{R}$, with the inner product $(f,g) = (f,g)_X = \int_0^1 f(x)g(x)dx$. Start with the functions

$$v_1(x) = 1, \quad v_2(x) = x, \quad v_3(x) = x^3$$

(a) Make these functions orthogonal, using the Gram-Schmidt process for these functions, with same ordering as in the previous exercise, to find the orthogonalized functions u_1, u_2, u_3.

(b) Now, assuming u_1, u_2, u_3 are orthogonal, given a function y, consider the functional regression problem

$$\min_w \|h(x,w) - y(x)\|_X^2$$

where $h(x,w) = w_1 u_1(x) - w_2 u_2(x) + w_3 u_3(x)$. Express the coefficients, w_1, w_2, w_3 of the minimizer, $h(x,w)$, in terms of inner products of y and the functions u_i. For $y(x) = \exp(5x)$, find w_1.

Exercise 5.27. Let $S = \{(x_1,y_1), \dots (x_m,y_m)\}$ where $X = [x_1, \dots, x_m]^\top = [0.01, 0.02, \dots, 0.99, 1.00]^\top$ is $m = 100$ equally spaced points in $[0,1]$. Let $y = \exp(5x)$ and let $y_i = y(x_i)$.

(a) Solve the feature regression problem on S_x with data Y using features $f(x) = [1, x, x^3]$ and with $h_1(x,w) = w \cdot f(x)$.

(b) Same problem, but find $h_2(x,v) = v \cdot g(x)$ where $g(x) = [u_1(x), u_2(x), u_3(x)]$, and u_i are the orthogonal features from the previous problem. Is $v = w$? Is $h_2 = h_1$?

(c) Now let $h_3(x,w)$ be the solution of the function regression problem with $X = [0,1]$. Approximate it by taking $m_2 = 1,000$ and solving the regression problem on the larger dataset. (You could also find the exact solution using integration/computer algebra). Plot $e(x) = h_1(x) - h_3(x)$, and find the mean squared error,

$$E(h_1, h_3, X) = \frac{1}{m}\sum_{i=1}^{m}(h_1(x_i) - h_3(x_i))^2$$

What is the value of the mean squared error? How does it compare to m (e.g. $1/m$, $1/m^2$)?