Exercise Bank PR25: Midterm 1

The following questions serve as additional practice material for the first midterm exam. The problems vary in difficulty and solutions are not provided. Each exercise is accompanied by its source, indicated in square brackets according to the following key:

- **[MML]** *Machine Learning: A Probabilistic Perspective* (2012) by Kevin Murphy
- **[ESL]** *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman
- **[PPA]** *Patterns, Predictions, and Actions* by Moritz Hardt and Ben Recht
- **[FPS]** *Foundations of Signal Processing* by Vetterli, Kovačević, and Goyal
- **[CVAA]** *Computer Vision: Algorithms and Applications*, 2nd Edition, by Richard Szeliski
- **[ISL]** *An Introduction to Statistical Learning with Applications in Python* by James, Witten, Hastie, Tibshirani, and Taylor

KNN Classifier

4.8.4: Curse of dimensionality of KNN and other local approaches [ISL]

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other *local* approaches that perform prediction using only observations that are *near* the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

4.8.8: Logistic regression vs. 1-nearest neighbors [ISL]

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Bayesian and Probabilistic Inference

2.9: Bayes rule for medical diagnosis [MML]

(Source: Koller.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people.

What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

2.2: Legal reasoning using Bayes rule [MML]

(Source: Peter Lee.) Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

a. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he is guilty." This is known as the *prosecutor's fallacy*. What is wrong with this argument?

b. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance." This is known as the *defender's fallacy*. What is wrong with this argument?

2.11: Normalization constant for a 1D Gaussian [MML]

The normalization constant for a zero-mean Gaussian is given by

$$Z = \int_a^b \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

where $a = -\infty$ and $b = \infty$. To compute this, consider its square

$$Z^2 = \int_a^b \int_a^b \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) dx\, dy.$$

Let us change variables from cartesian (x, y) to polar (r, θ) using $x = r\cos\theta$ and $y = r\sin\theta$. Since $dx\, dy = r\, dr\, d\theta$, and $\cos^2\theta + \sin^2\theta = 1$, we have

$$Z^2 = \int_0^{2\pi} \int_0^{\infty} r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr\, d\theta.$$

Evaluate this integral and hence show

$$Z = \sigma\sqrt{2\pi}.$$

Hint 1: Separate the integral into a product of two terms, the first of which (involving $d\theta$) is constant, so is easy.
Hint 2: If $u = e^{-r^2/2\sigma^2}$ then $\dfrac{du}{dr} = -\frac{1}{\sigma^2} r e^{-r^2/2\sigma^2}$, so the second integral is also easy (since $\int u'(r)\, dr = u(r)$).

1.1: Decision rule that minimizes error probability [PPA]

Let Y be a continuous random variable distributed over the closed interval $[0, 1]$. Under the null hypothesis H_0, Y is uniform:

$$p_{Y|H}(Y|H_0) = \begin{cases} 1 & 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Under the alternative hypothesis H_1, the conditional pdf of Y is as follows:

$$p_{Y|H}(Y|H_1) = \begin{cases} 2y & 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The *a priori* probability that Y is uniformly distributed is p.

1. Find the decision rule that minimizes the probability of error.
2. Find the closed form expression for the operating characteristic of the likelihood ratio test (LRT), i.e., P_D as a function of P_F for the LRT.
3. Suppose we require P_D to be at least $(1 + \epsilon)P_F$, where $\epsilon > 0$ is a fixed constant. Find $P_D^{\max}(\epsilon)$, the maximal value of P_D that is achievable under this constraint.

1.5: Minimum probability of error [PPA]

Suppose we are deciding between two hypotheses $H \in \{H_0, H_1\}$ based on observation $y \in \mathcal{Y} \in \mathbb{R}^+$. The models under the two hypotheses are

$$H_0 :\ p_{Y|H}(y|H_0) = \begin{cases} e^{-y} & y \ge 0, \\ 0 & \text{otherwise,} \end{cases} \qquad H_1 :\ p_{Y|H}(y|H_1) = \begin{cases} 2e^{-2y} & y \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

The prior beliefs are $\mathbb{P}(H = H_1) = p$ and $\mathbb{P}(H = H_0) = 1 - p$. Associated with the possible decisions are the costs $C_{00} = C_{11} = 0$, and $0 \le C_{01}, C_{10} \le \infty$, where C_{ij} is the cost of deciding $\hat{H}(y) = H_i$ when the correct hypothesis is $H = H_j$.

1. The decision rule $\hat{H}(\cdot)$ that minimizes the expected cost takes the form:

$$\hat{H} = \begin{cases} H_0 & \text{if } y \ge \gamma, \\ H_1 & \text{if } y < \gamma, \end{cases}$$

   Express γ in terms of C_{10}, C_{01} and p.

2. Express P_D as a function of P_F. Note that P_D and P_F are defined as:

$$P_D = \mathbb{P}(\hat{H} = H_1 \mid H = H_1), \qquad P_F = \mathbb{P}(\hat{H} = H_1 \mid H = H_0).$$

3. In the remainder of the problem, consider minimizing expected cost over a "3-way" decision rule, whereby, in addition to $\hat{H}(y) = H_0$ or $\hat{H}(y) = H_1$, one can decide $\hat{H}(y) =?$ ("I don't know") for some value(s) of y. Let us denote the corresponding costs using $C_{?0}$ and $C_{?1}$ when the underlying hypotheses are H_0 and H_1, respectively. Assume the costs are chosen to satisfy $0 = C_{00} \leq C_{?0} \leq C_{10}$ and $0 = C_{11} \leq C_{?1} \leq C_{01}$, so that admitting "I don't know" is less costly than making a wrong decision but more costly than making a correct decision.

   The optimal decision rule $\hat{H}_{\text{3-way}}(\cdot)$ in this case can be expressed in the form

$$\hat{H}_{\text{3-way}}(y) = \begin{cases} H_0 & \text{if } r(y) \leq u \text{ and } r(y) \leq v, \\ H_1 & \text{if } r(y) \geq u \text{ and } r(y) \geq v, \\ ? & \text{if } r(y) \geq v \text{ and } r(y) \leq w, \end{cases}$$

   where $r(y) = \dfrac{\pi_1(y)}{\pi_0(y)}$ with $\pi_0(y) = \mathbb{P}(H = H_0 | Y = y)$ and $\pi_1(y) = \mathbb{P}(H = H_1 | Y = y)$, and u, v, w are constants. Express u, v, w in terms of the costs $C_{ij}, i \in \{0, 1, ?\}$ and $j \in \{0, 1\}$.

4. Determine whether the following (*italicized*) statement is true or false, and justify your answer:

   For at least some value(s) of P_F, the optimal 3-way decision rule can achieve a greater P_D than that corresponding to the operating characteristic you found in part (b).

Linear Regression

3.7.5: Linear regression [ISL]

Consider the fitted values that result from performing linear regression without an intercept (bias). In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta},$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}.$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{ii'} y_{i'}.$$

What is $a_{ii'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

2.57: Least-squares solution to a system of linear equations [FPS]

For $\hat{y} = A\hat{x}$ and $\|y - \hat{y}\|_2$, show the following:

(i) Show that if y belongs to the column space of A, then $\hat{y} = y$.

(ii) Show that if y is orthogonal to the column space of A, then $\hat{y} = 0$.

(iii) Show that for the least-squares solution, the partial derivatives

$$\frac{\partial}{\partial \hat{x}_i} \|y - \hat{y}\|^2$$

are all zero.

2.60: Bayesian linear MMSE estimation via the orthogonality principle [FPS]

Let x and y be jointly distributed random vectors with $\mathbb{E}[x] = \mu_x$, $\mathbb{E}[y] = \mu_y$, and

$$\mathbb{E}\left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right] \left(\begin{bmatrix} x^* & y^* \end{bmatrix} - \begin{bmatrix} \mu_x^* & \mu_y^* \end{bmatrix}\right) = \begin{bmatrix} \Sigma_x & \Sigma_{x,y} \\ \Sigma_{x,y}^* & \Sigma_y \end{bmatrix}.$$

Use the projection theorem to find the LMMSE estimator of x as a function of y, that is, the optimal estimator of the form

$$\hat{x} = Ay + b.$$

3.1: Linear algebra problems for regression [PPA]

1. Let A be a $d \times n$ matrix. For any $\mu > 0$, show that

$$(AA^\top + \mu I)^{-1} A \;=\; A(A^\top A + \mu I)^{-1}.$$

2. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sequence of data points. Each y_i is a scalar and each x_i is a vector in $\mathbb{R}^d$. Let $X = [x_1, \ldots, x_n]^\top$ and $Y = [y_1, \ldots, y_n]^\top$. Consider the *regularized least squares problem*:

$$\min_{w \in \mathbb{R}^d} \; \|Xw - Y\|_2^2 + \mu \|w\|_2^2.$$

   Show that the optimum w_* is unique and can be written as the linear combination

$$w_* = \sum_{i=1}^{n} \alpha_i x_i$$

   for some scalars α. What are the coefficients α_i? *Hint: you may find eigendecomposition useful for representing α_i.*

Logistic Regression

4.8.1: Logistic regression [ISL]

Using a little bit of algebra, prove that

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

is equivalent to

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}.$$

In other words, the logistic function representation and the logit representation for the logistic regression model are equivalent.

10.10.2: Softmax function [ISL]

Consider the *softmax* function for modeling multinomial probabilities:

$$f_m(X) = \Pr(Y = m \mid X) = \frac{e^{Z_m}}{\sum_{\ell=0}^{9} e^{Z_\ell}},$$

and equivalently, for the logistic regression setting,

$$\Pr(Y = k \mid X = x) = \frac{\exp(\beta_{k0} + \beta_{k1} x_1 + \cdots + \beta_{kp} x_p)}{\sum_{\ell=1}^{K} \exp(\beta_{\ell 0} + \beta_{\ell 1} x_1 + \cdots + \beta_{\ell p} x_p)}.$$

(a) Show that if we add a constant c to each of the Z_ℓ, the probability $f_m(X)$ is unchanged.

(b) Show that if we add constants c_j, $j = 0, 1, \ldots, p$, to each of the corresponding coefficients β_{kj} for each of the classes, then the predictions at any new point x are unchanged.

This shows that the softmax function is *over-parameterized*. However, regularization and stochastic gradient descent (SGD) typically constrain the solutions so that this is not a problem.

Machine Learning

5.4: Activation and weight scaling [CVAA]

Consider the two hidden unit network shown in Figure (1), which uses ReLU activation functions and has no additive bias parameters. Your task is to find a set of weights that will fit the function

$$y = |x_1 + 1.1x_2|.$$

1. Can you guess a set of weights that will fit this function?
2. Starting with the weights shown in column (b), compute the activations for the hidden and final units as well as the regression loss for the nine input values $(x_1, x_2) \in \{-1, 0, 1\} \times \{-1, 0, 1\}$.
3. Now compute the gradients of the squared loss with respect to all six weights using the back-propagation chain rule equations and sum them up across the training samples to get a final gradient.
4. What step size should you take in the gradient direction, and what would your updated squared loss become?
5. Repeat this exercise for the initial weights in column (c) of Figure (1)
6. Given this new set of weights, how much worse is your error decrease, and how many iterations would you expect it to take to achieve a reasonable solution?
7. Would batch normalization help in this case?

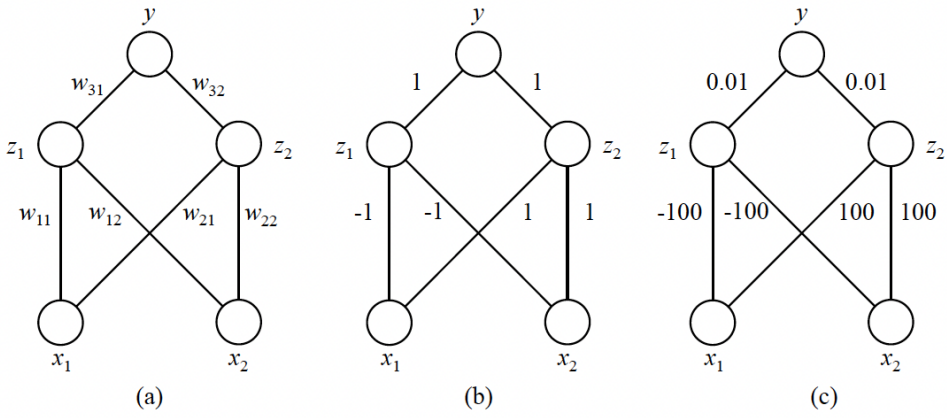

Figure 1: Simple two hidden unit network with a ReLU activation function and no bias parameters for regressing the function $y = |x_1 + 1.1x_2|$: (a) can you guess a set of weights that would fit this function?; (b) a reasonable set of starting weights; (c) a poorly scaled set of weights.

Code ex – 5.5: Function optimization with Newton [CVAA]

Consider the function

$$f(x, y) = x^2 + 20y^2$$

. Begin by solving for the following:

1. Calculate ∇f, i.e., the gradient of f.
2. Evaluate the gradient at $x = -20$, $y = 5$.

Implement some of the common gradient descent optimizers, which should take you from the starting point $x = -20$, $y = 5$ to near the minimum at $x = 0$, $y = 0$. Try each of the following optimizers:

1. Standard gradient descent.
2. Gradient descent with momentum, starting with the momentum term as $\rho = 0.99$.
3. Adam, starting with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Play around with the learning rate α. For each experiment, plot how x and y change over time.

How do the optimizers behave differently? Is there a single learning rate that makes all the optimizers converge towards $x = 0, y = 0$ in under 200 steps? Does each optimizer monotonically trend towards $x = 0, y = 0$?

10.10.1: Basic NN exercise [ISL]

Consider a neural network with two hidden layers: $p = 4$ input units, 2 units in the first hidden layer, 3 units in the second hidden layer, and a single output.

(a) Draw a picture of the network, similar to Figures 2.

(b) Write out an expression for $f(X)$, assuming ReLU activation functions. Be as explicit as you can!

(c) Now plug in some values for the coefficients and write out the value of $f(X)$.

(d) How many parameters are there?

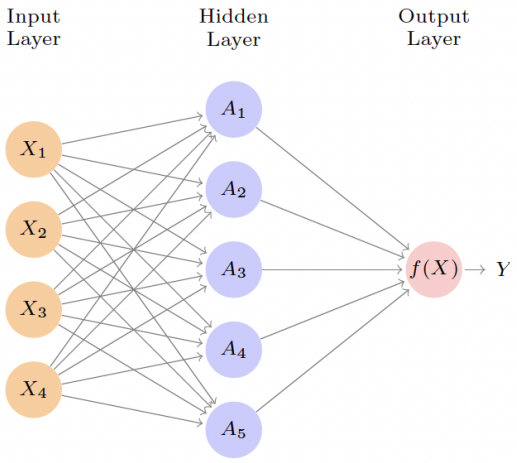

Figure 2: Neural network with 4 input units, 2 units in the first hidden layer, 3 units in the second hidden layer, and one output.