CSE 250B: Machine learning **Winter 2019**

Homework 1 — Nearest neighbor and statistical learning

1. *Risk of a random classifier.* A particular data set has 4 possible labels, with the following frequencies:

Label	Frequency
A	50%
B	20%
C	20%
D	10%

   (a) What is the error rate (risk) of a classifier that picks a label (A,B,C,D) uniformly at random?

   (b) One very simple type of classifier just returns the same label, always. What label should it return, and what will its error rate be?

2. *Discrete and continuous distributions.* In this class, we will deal with both discrete and continuous random variables. Let's look at examples of each.

   (a) A discrete random variable X is said to have Poisson distribution with parameter λ if it can take on values in $\{0, 1, 2, \ldots\}$, with

$$\Pr(X = i) \;=\; e^{-\lambda}\frac{\lambda^i}{i!}.$$

   You can check that these probabilities sum to 1 by looking at the Taylor series for e^λ. Can you give another example of a discrete distribution that assigns positive probabilities to infinitely many values?

   (b) A continuous random variable X has uniform distribution over $[a, b]$ if it has *density function*

$$f(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

   This means that the probability that X lies in some interval $[a', b'] \subseteq [a, b]$ is

$$\int_{a'}^{b'} f(x)\, dx.$$

   What is the probability that X is exactly $(a + b)/2$?

3. *Complexity analysis for k-d tree with defeatist search.* Suppose a k-d tree is built on n data points in $\mathbb{R}^d$, by splitting until each leaf node has $\le n_o$ points.

   (a) What is the time complexity of building the tree, as a function of n, d, and n_o? Justify your answer carefully.

   (b) What is the time complexity of answering a query using defeatist search?

4. *Properties of metrics.* Which of the following distance functions are metrics? In each case, either prove it is a metric or give a counterexample showing that is isn't.

   (a) $d_1 + d_2$, where d_1 and d_2 are each metrics.

   (b) Let's say Σ is a finite set and $\mathcal{X} = \Sigma^m$. The *Hamming distance* on $\mathcal{X}$ is

   $$d(x, y) = \#\text{ of positions on which } x \text{ and } y \text{ differ.}$$

   (c) Squared Euclidean distance on $\mathbb{R}^m$, that is,

   $$d(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2.$$

   (It might be easiest to consider the case $m = 1$.)

5. *A joint distribution over data and labels.* A distribution over two-dimensional data points $X = (X_1, X_2) \in \mathbb{R}^2$ and their labels $Y \in \{0, 1\}$ is specified as follows:

   - The two labels are equally likely, that is, $\Pr(Y = 0) = \Pr(Y = 1) = 1/2$.
   - When $Y = 0$, the points X are uniformly distributed in the square $[-2, -1] \times [-2, -1]$.
   - When $Y = 1$, the points X are uniformly distributed in the square $[1, 3] \times [2, 4]$.

   (a) In a two-dimensional plane, sketch the regions where points (x_1, x_2) might fall. Label one of these regions with $y = 0$ and the other with $y = 1$.

   (b) What is the marginal distribution of X_1? Specify it exactly.

   (c) What is the marginal distribution of X_2?

6. *Two ways of specifying a joint distribution over data and labels.* Consider the following distribution over two-dimensional data points $X = (X_1, X_2)$ and their labels $Y \in \{0, 1\}$:

   - $\Pr(Y = 1) = 1/4$
   - When $Y = 0$, points X are uniformly distributed in the rectangle $[0, 3] \times [0, 1]$.
   - When $Y = 1$, points X are uniformly distributed in the rectangle $[-1, 1] \times [0, 1]$.

   Rewrite this distribution in the form of two functions: μ, the density function for X; and η, the conditional distribution of Y given X.

7. *Bayes optimality.* Consider the following setup:

   - Input space $\mathcal{X} = [-1, 1] \subset \mathbb{R}$.
   - Input distribution: $\mu(x) = |x|$.
   - Label space $\mathcal{Y} = \{0, 1\}$.
   - Conditional probability function

   $$\eta(x) = \Pr(Y = 1 | X = x) = \begin{cases} 0.2 & \text{if } x < -0.5 \\ 0.8 & \text{if } -0.5 \leq x \leq 0.5 \\ 0.4 & \text{if } x > 0.5 \end{cases}$$

   (a) What is the Bayes optimal classifier in this setting? What is the optimal risk R^*?

(b) Suppose we obtain the following training set of four labeled points:

$$(-0.8, 0), (-0.4, 1), (0.2, 1), (0.8, 0).$$

What is the decision boundary of 1-NN using this training set? What is the (true) error rate of this classifier, on the underlying distribution given by μ and η?

(c) In a binary setting, there are two possible errors: $0 \to 1$ (label is 0 but prediction is 1) or $1 \to 0$ (label is 1 but prediction is 0). Suppose these errors have different costs, c_{01} and c_{10}, respectively. We can then define the *cost-sensitive risk* of a classifier $h : \mathcal{X} \to \{0, 1\}$ as

$$R(h) = c_{01}\Pr(Y = 0, h(X) = 1) + c_{10}\Pr(Y = 1, h(X) = 0).$$

In the example above, what is the classifier that minimizes this cost-sensitive risk, if $c_{01} = 1$ and $c_{10} = 0.1$?

(c) Now consider a setting with $\mathcal{Y} = \{0, 1\}$ and with arbitrary $\mathcal{X}, \mu, \eta, c_{01}, c_{10}$. Write down an expression for the classifier with minimum cost-sensitive risk.

CSE 250B: Homework 1 Solutions

1. *Risk of a random classifier.*

   (a) No matter what the correct label is, the probability that a random classifier selects it is 0.25. Therefore, this classifier has risk (error probability) 0.75.

   (b) We should return the label with the highest probability, which is A. The risk of this classifier is the probability that the label is something else, namely 0.5.

2. *Discrete and continuous distributions.*

   (a) Another example of a discrete distribution with infinite support is the *geometric distribution*. The simplest case of this has possible outcomes $0, 1, 2, \ldots$, where the probability of outcome i is $1/2^{i+1}$.

   (b) If X follows a uniform distribution over $[a, b]$ (where $a < b$), the probability that X takes on any specific value is 0.

3. *Complexity analysis for k-d tree with defeatist search.*

   (a) Let's assume that we are given the d-dimensional data points in the form of an $n \times d$ matrix. We will construct a tree data structure whose leaves each contain at most n_o of these data points (more precisely, a list of the row indices corresponding to the points).

   At each internal node of the tree, containing (say) m points:

   - The time taken to choose a coordinate for splitting is $O(md)$, if we pick the coordinate with highest variance.
   - We can use a linear-time median-finding algorithm to find the split point.
   - We partition the points into left and right groups, also in linear time.

   Therefore, the total time taken for this node is $O(md)$ and thus the time for constructing an entire level of the tree is $O(nd)$.

   There are n points in the data set and with each successive level, the number of points per cell is halved, until we reach leaf nodes with $\leq n_o$ points. So the height of the tree is $\log(n/n_o)$.
   Therefore, the total complexity of building a k-d tree as specified in the problem is $O(nd \log(n/n_o))$.

   (b) To answer a query, we first move to the appropriate leaf of the tree, which takes time $O(\log(n/n_o))$ (constant time per internal node along the way), and we then look for the nearest neighbor within that leaf, which takes time $O(n_o d)$. The total query time is thus $O(n_o d + \log(n/n_o))$.

4. *Properties of metrics.* Recall that d is a distance metric if and only if it satisfies the following properties:

   (P1) $d(x, y) \geq 0$

   (P2) $d(x, y) = 0 \iff x = y$

   (P3) $d(x, y) = d(y, x)$ (symmetry)

   (P4) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

   (a) If d_1 and d_2 are metrics, then so is $g(x, y) = d_1(x, y) + d_2(x, y)$. All four properties can be verified directly.

   (P1) $g(x, y) \geq 0$ because it is the sum of two nonnegative values.

   (P2) Pick any x, y.

$$\begin{aligned} g(x, y) = 0 &\iff d_1(x, y) + d_2(x, y) = 0 \\ &\iff d_1(x, y) = 0 \text{ and } d_2(x, y) = 0 \text{ (since both nonnegative)} \\ &\iff x = y \end{aligned}$$

(P3) $g(x,y) = d_1(x,y) + d_2(x,y) = d_1(y,x) + d_2(y,x) = g(y,x)$.

(P4) For any x, y, z,

$$\begin{aligned} g(x,z) &= d_1(x,z) + d_2(x,z) \\ &\leq (d_1(x,y) + d_1(y,z)) + (d_2(x,y) + d_2(y,z)) \\ &= (d_1(x,y) + d_2(x,y)) + (d_1(y,z) + d_2(y,z)) \\ &= g(x,y) + g(y,z) \end{aligned}$$

(b) Hamming distance is a metric.

(P1) $d(x,y) \geq 0$ because number of positions at which two strings differ can't be negative.

(P2) $d(x,x) = 0$ because a string differs from itself at no positions. Also, if $x \neq y$, there will be at least one position where x and y differ and hence $d(x,y) \geq 0$.

(P3) $d(x,y) = d(y,x)$ because x differs from y at exactly the same positions where y differs from x.

(P4) Pick any $x, y, z \in \Sigma^m$. Let A denote the positions at which x, y differ: $A = \{i : x_i \neq y_i\}$, so that $d(x,y) = |A|$. Likewise, let B be the positions at which y, z differ and let C be the positions at which x, z differ.
Now, if $x_i = y_i$ and $y_i = z_i$, then $x_i = z_i$. Thus $C \subseteq A \cup B$, whereupon $d(x,z) = |C| \leq |A| + |B| = d(x,y) + d(y,z)$.

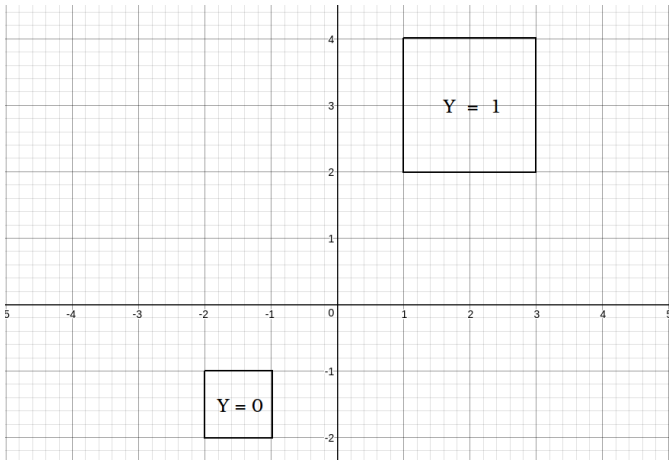
(c) Squared Euclidean distance is not a metric as it does not satisfy the triangle inequality. Consider the following three points in $\mathbb{R}$: $x = 1, y = 4, z = 5$.

$$\begin{aligned} d(x,z) &= (1-5)^2 = 16 \\ d(x,y) &= (1-4)^2 = 9 \\ d(y,z) &= (4-5)^2 = 11 \end{aligned}$$

Here $d(x,z) > d(x,y) + d(y,z)$.

5. *A joint distribution over data and labels.*

(a) Graph with regions where (x_1, x_2) might fall.



(b) Let μ_1 denote the density function of X_1.

$$\mu_1(x_1) = \begin{cases} 1/2 & \text{if } -2 \leq x_1 \leq -1 \\ 1/4 & \text{if } 1 \leq x_1 \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

(c) Let μ_2 denote the density function of X_2.

$$\mu_2(x_2) = \begin{cases} 1/2 & \text{if } -2 \le x_2 \le -1 \\ 1/4 & \text{if } 2 \le x_2 \le 4 \\ 0 & \text{elsewhere} \end{cases}$$

6. *Two ways of specifying a joint distribution over data and labels.*

The marginal distribution of $x = (x_1, x_2)$ is given by the following density function:

$$\mu(x_1, x_2) = \begin{cases} 1/8 & \text{if } -1 \le x_1 < 0 \\ 3/8 & \text{if } 0 \le x_1 < 1 \\ 1/4 & \text{if } 1 \le x_1 \le 3 \end{cases}$$

The conditional distribution of y given $x = (x_1, x_2)$ is

$$\eta(x) \;=\; \Pr(Y = 1 | X = (x_1, x_2)) \;=\; \begin{cases} 1 & \text{if } -1 \le x_1 < 0 \\ 1/3 & \text{if } 0 \le x_1 < 1 \\ 0 & \text{if } 1 \le x_1 \le 3 \end{cases}$$

7. *Bayes optimality.*

(a) The Bayes-optimal classifier predicts 1 when $-0.5 \le x \le 0.5$, and 0 elsewhere. Its risk (probability of being wrong) is:

$$R^* \;=\; \int_{-1}^{1} \min(\eta(x), 1 - \eta(x))\, \mu(x)\, dx \;=\; \int_{-1}^{0.5} 0.2|x|\, dx \;+\; \int_{0.5}^{1} 0.4|x|\, dx \;=\; 0.275.$$

(b) The 1-NN classifier based on the four given points predicts as follows:

$$h(x) \;=\; \begin{cases} 1 & \text{if } -0.6 \le x \le 0.5 \\ 0 & \text{if } x < -0.6 \text{ or } x > 0.5 \end{cases}$$

Notice that this differs slightly from the Bayes optimal classifier. The risk of rule h is

$$\begin{aligned} R(h) &= \int_{-1}^{1} \Pr(y \ne h(x) \mid x)\, \mu(x)\, dx \\ &= \int_{-1}^{-0.6} 0.2|x|\, dx + \int_{-0.6}^{-0.5} 0.8|x|\, dx + \int_{-0.5}^{0.5} 0.2|x|\, dx + \int_{0.5}^{1} 0.4|x|\, dx \;=\; 0.308. \end{aligned}$$

(c) The cost of predicting 1 when the true label is 0 is ten times the cost of predicting 0 when the true label is 1. The best thing to do is to simply predict 0 everywhere.

(d) The classifier with smallest cost-sensitive risk is:

$$h^*(x) = \begin{cases} 1 & \text{if } c_{01}(1 - \eta(x)) \le c_{10}\eta(x) \\ 0 & \text{if } c_{01}(1 - \eta(x)) > c_{10}\eta(x) \end{cases}$$

CSE 250B: Machine learning **Winter 2018**

Homework 2 — Statistical learning and Linear Algebra

1. *Error rate of 1-NN classifier.*

   (a) Give an example of a data set with just three points (x, y) for which the 1-NN classifier does *not* have zero training error (that is, it makes mistakes on the training set).

   (b) Is 1-NN classification necessarily consistent in cases where the Bayes risk R^* is zero?

2. *Bayes optimality in a multi-class setting.* In lecture, we discussed the setup of statistical learning theory in binary classification. We will now generalize this to situations in which the label space $\mathcal{Y}$ is possibly larger, though still finite.

   Suppose $|\mathcal{Y}| = \{1, 2, \ldots, \ell\}$, where $\ell > 2$. We will replace our earlier conditional probability function η by a set of ℓ such functions, denoted $\eta_1, \ldots, \eta_\ell$. Each η_i is a function from $\mathcal{X}$ to $[0, 1]$ and has the following meaning:

   $$\eta_i(x) = \Pr(Y = i | X = x).$$

   In particular, therefore, $\sum_i \eta_i(x) = 1$ for any x.

   What is the Bayes-optimal classifier – that is, the classifier with minimum error – in this case? Specify it precisely, in terms of the η functions.

3. *Classification with an abstain option.* As usual, we can factor a distribution over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{0, 1\}$, into a marginal distribution μ on $\mathcal{X}$ and a conditional probability function $\eta(x) = \Pr(Y = 1 | X = x)$.

   In some situations, it is useful to allow a classifier to *abstain* from predicting on instances x on which it is unsure. Such instances can then be treated separately. Suppose the cost structure is set up so that:

   - If the classifier makes a prediction (either 0 or 1), it incurs no cost if the prediction is correct and a cost of 1 if the prediction is wrong.
   - If the classifier abstains, it incurs a fixed cost θ, which is some real number between 0 and 1/2.

   What classifier $h : \mathcal{X} \to \{0, 1, \text{abstain}\}$ has minimum expected cost? You should write $h(x)$ as a function of $\eta(x)$ and θ.

4. *The statistical learning assumption.* In each of the following cases, say whether or not you feel the statistical learning assumption would hold. If not, explain the nature of the violation (for instance, μ is changing but not η, or η is changing, or the sampling is not independent and random). The answers may be subjective, so explain your position carefully.

   (a) A music studio wants to build a classifier that predicts whether a proposed song will be a commercial success. It builds a data set of all the songs it has considered in the past, labeled according to whether or not that song was a hit; and it uses this data to train a classifier.

   (b) A bank wants to build a classifier that predicts whether a loan applicant will default or not. It builds a data set based on all loans it accepted over the past ten years, labeled according to whether or not they went into default. These are then used to train the classifier.

(c) An online dating site uses machine learning prediction techniques to decide whether a pair of people are likely to be compatible with each other. Their classifier has worked well on the west coast, and now they decide to take it to the national level.

5. *Conditional probability.* A particular child is always in one of two possible moods: `happy` and `sad`. The prior probabilities of these are:

$$\pi(\texttt{happy}) = \frac{3}{4}, \quad \pi(\texttt{sad}) = \frac{1}{4}.$$

One can usually judge his mood by how talkative he is. After much observation, it has been determined that:

- When he is happy,

$$\Pr(\text{talks a lot}) = \frac{2}{3}, \quad \Pr(\text{talks a little}) = \frac{1}{6}, \quad \Pr(\text{completely silent}) = \frac{1}{6}$$

- When he is sad,

$$\Pr(\text{talks a lot}) = \frac{1}{6}, \quad \Pr(\text{talks a little}) = \frac{1}{6}, \quad \Pr(\text{completely silent}) = \frac{2}{3}$$

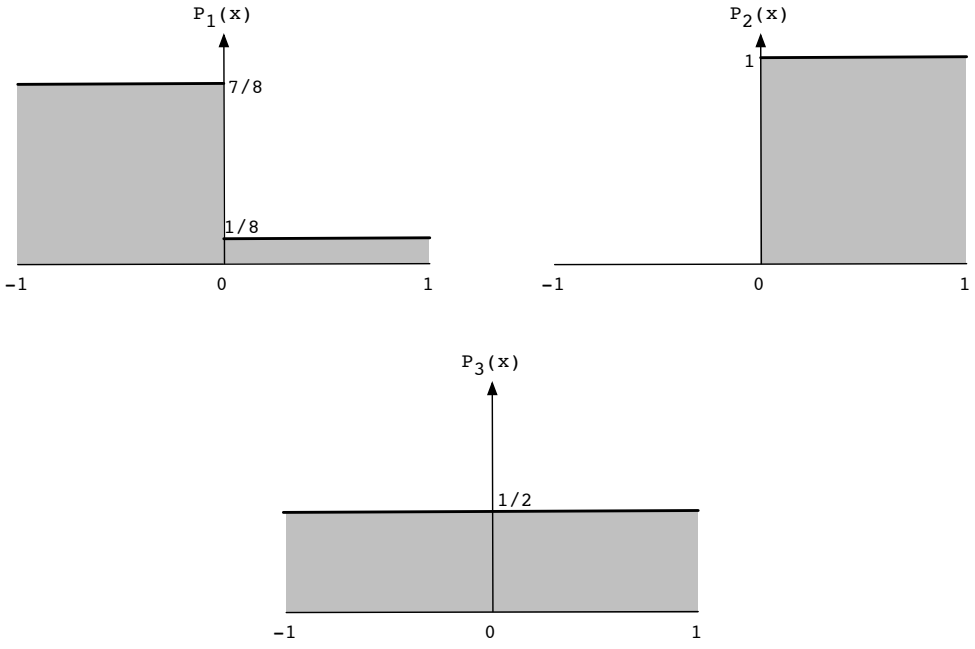
(a) Tonight, the child is just talking a little. What is his most likely mood?

(b) What is the probability of the prediction in part (a) being incorrect?

6. *Bayes optimal classifier.* Suppose $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = \{1, 2, 3\}$, and that the individual classes have weights

$$\pi_1 = 1/3, \pi_2 = 1/6, \pi_3 = 1/2$$

and densities P_1, P_2, P_3 as shown below.



(a) What is the best (Bayes-optimal) classifier h^*? Specify it exactly, as a function from $\mathcal{X}$ to $\mathcal{Y}$.

(b) What is the error rate of h^*?

7. Find the unit vector in the same direction as $x = (1, 2, 3)$.
8. Find all unit vectors in $\mathbb{R}^2$ that are orthogonal to $(1, 1)$.
9. How would you describe the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$?
10. The function $f(x) = 2x_1 - x_2 + 6x_3$ can be written as $w \cdot x$ for $x \in \mathbb{R}^3$. What is w?
11. For a certain pair of matrices A, B, the product AB has dimension 10×20. If A has 30 columns, what are the dimensions of A and B?
12. We have n data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ and we store them in a matrix X, one point per row.

    (a) What is the dimension of X?

    (b) What is the dimension of XX^T?

    (c) What is the (i, j) entry of XX^T, simply?
13. Vector x has length 10. What is $x^T x x^T x x^T x$?
14. For $x = (1, 3, 5)$ compute $x^T x$ and xx^T.
15. Vectors $x, y \in \mathbb{R}^d$ both have length 2. If $x^T y = 2$, what is the angle between x and y?
16. The quadratic function $f : \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x) = 3x_1^2 + 2x_1x_2 - 4x_1x_3 + 6x_3^2$$

    can be written in the form $x^T M x$ for some **symmetric** matrix M. What is M?
17. Which of the following matrices is necessarily symmetric?

    (a) AA^T for arbitrary matrix A.

    (b) $A^T A$ for arbitrary matrix A.

    (c) $A + A^T$ for arbitrary square matrix A.

    (d) $A - A^T$ for arbitrary square matrix A.
18. Let $A = \text{diag}(1, 2, 3, 4, 5, 6, 7, 8)$.

    (a) What is $|A|$?

    (b) What is A^{-1}?
19. Vectors $u_1, \ldots, u_d \in \mathbb{R}^d$ all have unit length and are orthogonal to each other. Let U be the $d \times d$ matrix whose rows are the u_i.

    (a) What is UU^T?

    (b) What is U^{-1}?
20. Matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & z \end{pmatrix}$ is singular. What is z?

CSE 250B: Homework 2 Solutions

1. *Error rate of 1-NN classifier.*

   (a) Consider a training set in which the same point x appears twice, but with different labels. The training error of 1-NN on this data will not be zero.

   (b) We mentioned in class that the risk of the 1-NN classifier, $R(h_n)$, approaches $2R^*(1-R^*)$ as $n \to \infty$ where R^* is the Bayes risk. If $R^* = 0$, this means that the 1-NN classifier is consistent: $R(h_n) \to 0$.

2. *Bayes optimality in a multi-class setting.* The Bayes-optimal classifier predicts the label that is most likely:

$$h^*(x) = \arg\max_{i \in |\mathcal{Y}|} \eta_i(x)$$

3. *Classification with an abstain option.* The classifier should abstain whenever the probability of error exceeds θ:

$$h^*(x) = \begin{cases} \text{abstain} & \text{if } \theta < \eta(x) < 1-\theta \\ 1 & \text{if } \eta(x) \geq 1-\theta \\ 0 & \text{if } \eta(x) \leq \theta \end{cases}$$

4. *The statistical learning assumption.*

   (a) Here, μ is the distribution over proposed songs, while η tells us which songs will be successful. Both are likely to change with time, violating the statistical learning assumption. However, the drift might be quite slow, so a classifier trained today may work well for another year or two before needing to be re-trained.

   (b) In this example, the bank's data set consists only of loans it *accepted*. It is not a random sample from μ, which is the distribution over all loan applications. This is a severe violation of the i.i.d. sampling requirement.

   (c) The move from the west coast to the entire country means that μ is changing, and it is possible that η is changing as well. Technically, this violates the statistical learning assumption; but it is possible that the change in distribution may not be very severe.

5. *Conditional probability.*

   (a) He is most likely to be in `happy` mood.

   (b) The probability of the baby being happy is Pr(`happy`|talks a little).

$$\begin{aligned} \Pr(\texttt{happy}|\text{talks a little}) &= \frac{\Pr(\text{talks a little}|\texttt{happy})\Pr(\texttt{happy})}{\Pr(\text{talks a little})} \\ &= \frac{\Pr(\text{talks a little}|\texttt{happy})\Pr(\texttt{happy})}{\Pr(\text{talks a little}|\texttt{happy})\Pr(\texttt{happy}) + \Pr(\text{talks a little}|\texttt{sad})\Pr(\texttt{sad})} \\ &= \frac{\frac{1}{6} \times \frac{3}{4}}{\frac{1}{6} \times \frac{3}{4} + \frac{1}{6} \times \frac{1}{4}} \;=\; \frac{3}{4} \end{aligned}$$

   Therefore, the probability of the prediction begin wrong is $1 - \Pr(\texttt{happy}|\text{talks a little}) = 1/4$.

6. *Bayes optimal classifier.*

   (a)

$$h^*(x) = \arg\min_{i \in \mathcal{Y}} \pi_i P_i(x) = \begin{cases} 1 & \text{if } -1 \leq x \leq 0 \\ 3 & \text{if } 0 < x \leq 1 \end{cases}$$

(b) The probability density function of $\mathcal{X}$ is

$$\mu(x) = \begin{cases} \frac{13}{24} & x \in [-1, 0] \\ \frac{11}{24} & x \in (0, 1] \end{cases}$$

Looking at all the ways to be wrong, the error rate is

$$\Pr(y = 1 \text{ and } x > 0) + \Pr(y = 2) + \Pr(y = 3 \text{ and } x \leq 0) \;=\; \frac{1}{3} \cdot \frac{1}{8} + \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{2} \;=\; \frac{11}{24}$$

7. $(1/\sqrt{14}, 2/\sqrt{14}, 3/\sqrt{14})$
8. $(-1/\sqrt{2}, 1/\sqrt{2})$ and $(1/\sqrt{2}, -1/\sqrt{2})$
9. $x \cdot x = 25 \Leftrightarrow \|x\| = 5$. All points of length 5: a sphere, centered at the origin, of radius 5.
10. $f(x) = 2x_1 - x_2 + 6x_3 = w \cdot x$ for $w = (2, -1, 6)$.
11. A is 10×30 and B is 30×20
12. (a) X is $n \times d$
    (b) XX^T is $n \times n$
    (c) $(XX^T)_{ij} = x^{(i)} \cdot x^{(j)}$
13. $((x^T x)(x^T x)(x^T x)) = (\|x\|^2)^3 = 10^6$
14. $x^T x = \|x\|^2 = 35$ and

$$xx^T = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 9 & 15 \\ 5 & 15 & 25 \end{bmatrix}$$

15. The angle θ between x and y satisfies $\cos\theta = x^T y / \|x\|\|y\| = 1/2$, so θ is 60 degrees.
16.

$$M = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix}$$

17. *Symmetric Matrices*
    (a) $(AA^T)^T = (A^T)^T A^T = AA^T$, Thus AA^T is symmetric.
    (b) $(A^T A)^T = A^T (A^T)^T = A^T A$, Thus $A^T A$ is symmetric.
    (c) $(A + A^T)^T = (A^T + A) = (A + A^T)$, Thus $(A + A^T)$ is symmetric
    (d) $(A - A^T)^T = (A^T - A) \neq (A - A^T)$, Thus $(A - A^T)$ need not be symmetric
18. (a) $|A| = 8! = 40320$
    (b) $A^{-1} = \text{diag}(1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8)$
19. *Orthonormal matrices*
    (a) UU^T is the identity matrix
    (b) $U^{-1} = U^T$
20. Since A is singular matrix, $|A| = 0 \implies z - 6 = 0 \implies z = 6$

CSE 250B: Machine learning **Winter 2019**

Homework 3 — Regression, logistic regression, unconstrained optimization

1. *Example of regression with one predictor variable.* Consider the following simple data set of four points (x, y):

$$(1, 1), (1, 3), (4, 4), (4, 6).$$

   (a) Suppose you had to predict y without knowledge of x. What value would you predict? What would be its mean squared error (MSE) on these four points?

   (b) Now let's say you want to predict y based on x. What is the MSE of the linear function $y = x$ on these four points?

   (c) Find the line $y = ax + b$ that minimizes the MSE on these points. What is its MSE?

2. *Lines through the origin.* Suppose that we have data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)}, y^{(i)} \in \mathbb{R}$, and that we want to fit them with a line that passes through the origin. The general form of such a line is $y = ax$: that is, the sole parameter is $a \in \mathbb{R}$.

   (a) The goal is to find the value of a that minimizes the squared error on the data. Write down the corresponding loss function.

   (b) Using calculus, find the optimal setting of a.

3. Suppose that $y = x_1 + x_2 + \cdots + x_{10}$, where:

   - $x_1, \ldots, x_{10}$ are independent, and
   - the x_i each have a Gaussian distribution with mean 1 and variance 1.

   (a) We wish to express y as a linear function of just $x_1, \ldots, x_5$. What is the linear function that minimizes MSE?

   (b) What is the mean squared error of the function in (a)?

4. We have a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. We want to express y as a linear function of x, but the error penalty we have in mind is not the usual squared loss: if we predict $\widehat{y}$ and the true value is y, then the penalty should be the absolute difference, $|y - \widehat{y}|$. Write down the loss function that corresponds to the total penalty on the training set.

5. We have n data points in $\mathbb{R}^d$ and we want to compute all pairwise dot products between them. Show that this can be achieved by a *single* matrix multiplication.

6. *Discovering relevant features in regression.* The data file `mystery.dat` contains pairs (x, y), where $x \in \mathbb{R}^{100}$ and $y \in \mathbb{R}$. There is one data point per line, with comma-separated values; the very last number in each line is the y-value.

   In this data set, y is a linear function of just *ten* of the features in x, plus some noise. Your job is to identify these ten features.

(a) Explain your strategy in one or two sentences.

(b) Which ten features did you identify? You need only give their coordinate numbers, from 1 to 100.

7. A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. Write down a precise expression for the set of points x with

   (a) $\Pr(y = 1|x) = 1/2$

   (b) $\Pr(y = 1|x) = 3/4$

   (c) $\Pr(y = 1|x) = 1/4$

8. Suppose that in a bag-of-words representation, we decide to use the following vocabulary of five words: (`is`, `flower`, `rose`, `a`, `an`). What is the vector form of the sentence "A rose is a rose is a rose"?

9. We are given a set of data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, and we want to find a single point $z \in \mathbb{R}^d$ that minimizes the loss function

$$L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2.$$

   Use calculus to determine z, in terms of the $x^{(i)}$.

10. Consider the following loss function on vectors $w \in \mathbb{R}^4$:

$$L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4.$$

    (a) What is $\nabla L(w)$?

    (b) Suppose we use gradient descent to minimize this function, and that the current estimate is $w = (0, 0, 0, 0)$. If the step size is η, what is the next estimate?

    (c) What is the minimum value of $L(w)$?

    (d) Is there is a unique solution w at which this minimum is realized?

11. Consider the loss function for ridge regression (ignoring the intercept term):

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 \; + \; \lambda \|w\|^2$$

    where $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$ are the data points and $w \in \mathbb{R}^d$. There is a closed-form equation for the optimal w (as we saw in class), but suppose that we decide instead to minimize the function using local search.

    (a) What is $\nabla L(w)$?

    (b) Write down the update step for gradient descent.

    (c) Write down a stochastic gradient descent algorithm.

CSE 250B: Homework 3 Solutions

1. *Regression with one predictor variable*

   (a) We will predict the mean of the y-values: $\widehat{y} = (1+3+4+6)/4 = 3.5$. The MSE of this prediction is exactly the variance of the y-values, namely:

   $$\text{MSE} = \frac{(1-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (6-3.5)^2}{4} = 3.25.$$

   (b) If we simply predict x, the MSE is

   $$\frac{1}{4}\sum_{i=1}^{4}(y^{(i)} - x^{(i)})^2 = \frac{1}{4}\left((1-1)^2 + (1-3)^2 + (4-4)^2 + (4-6)^2\right) = 2.$$

   (c) We saw in class that the MSE is minimized by choosing

   $$a = \frac{\sum_i (y^{(i)} - \overline{y})(x^{(i)} - \overline{x})}{\sum_i (x^{(i)} - \overline{x})^2}$$
   $$b = \overline{y} - a\overline{x}$$

   where $\overline{x}$ and $\overline{y}$ are the mean values of x and y, respectively. This works out to $a = 1, b = 1$; and thus the prediction on x is simply $x + 1$. The MSE of this predictor is:

   $$\frac{1}{4}\left(1^2 + 1^2 + 1^2 + 1^2\right) = 1.$$

2. *Lines through the origin*

   (a) The loss function is

   $$L(a) = \sum_{i=1}^{n}(y^{(i)} - ax^{(i)})^2$$

   (b) The derivative of this function is:

   $$\frac{dL}{da} = -2\sum_{i=1}^{n}(y^{(i)} - ax^{(i)})x^{(i)}.$$

   Setting this to zero yields

   $$a = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n} {x^{(i)}}^2}.$$

3. (a) The best predictor is $\hat{y} = x_1 + x_2 + x_3 + x_4 + x_5 + 5$: to minimize the fluctuations due to $x_6 + \cdots + x_{10}$, we use its mean.

   (b) All errors come from the variance in $x_6 + \cdots + x_{10}$, so
   $\text{MSE} = \text{var}(x_6 + \cdots + x_{10}) = \text{var}(x_6) + \cdots + \text{var}(x_{10}) = 5$.

4. The loss induced by a linear predictor $w \cdot x + b$ is

   $$L(w, b) = \sum_{i=1}^{n} |y^{(i)} - (w \cdot x^{(i)} + b)|.$$

5. Define

$$X = \begin{bmatrix} \leftarrow x^{(1)} \rightarrow \\ \leftarrow x^{(2)} \rightarrow \\ \vdots \\ \leftarrow x^{(n)} \rightarrow \end{bmatrix}$$

$$XX^T = \begin{bmatrix} x^{(1)} \cdot x^{(1)} & x^{(1)} \cdot x^{(2)} & \cdots & x^{(1)} \cdot x^{(n)} \\ x^{(2)} \cdot x^{(1)} & x^{(2)} \cdot x^{(2)} & \cdots & x^{(2)} \cdot x^{(n)} \\ x^{(n)} \cdot x^{(1)} & x^{(n)} \cdot x^{(2)} & \cdots & x^{(n)} \cdot x^{(n)} \end{bmatrix}$$

6. *Discovering relevant features in regression.*

   (a) A sensible strategy is to do linear regression using the Lasso, and to choose a regularization constant λ that yields roughly 10 non-zero coefficients.

   (b) The smallest value of λ we tried that gave nonzero coefficients for 10 features is 0.4. This yielded the following features (numbering starting at 1): $2, 3, 5, 7, 11, 13, 17, 19, 23, 29$.

7. *Logistic regression.* Since

$$\Pr(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}},$$

   we can rearrange terms to get

$$w \cdot x + b = \ln \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)}$$

   (a) $w \cdot x + b = \ln 1 = 0$

   (b) $w \cdot x + b = \ln 3$

   (c) $w \cdot x + b = -\ln 3$

8. With vocabulary $V = \{is, flower, rose, a, an\}$, the bag-of-words representation of the sentence "a rose is a rose is a rose" is $(2, 0, 3, 3, 0)$.

9. We want to find the $z \in \mathbb{R}^d$ that minimizes

$$L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_j^{(i)} - z_j)^2.$$

   Taking partial derivatives, we have

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{n} -2(x_j^{(i)} - z_j) = 2nz_j - 2\sum_{i=1}^{n} x_j^{(i)}.$$

   Thus

$$\nabla L(z) = 2nz - 2\sum_{i=1}^{n} x^{(i)}.$$

   Setting $\nabla L(z) = 0$ and solving for z, gives us

$$z^* = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}.$$

10. $L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4$

(a) The derivative is

$$\nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4)$$

(b) The derivative at $w = (0, 0, 0, 0)$ is $(2, -4, 0, 0)$. Thus the update at this point is:

$$w_{new} = w - \eta \nabla L(w) = (0, 0, 0, 0) - \eta(2, -4, 0, 0) = (-2\eta, 4\eta, 0, 0).$$

(c) To find the minimum value of $L(w)$, we will equate $\nabla L(w)$ to zero:

- $2w_1 + 2 = 0 \implies w_1 = -1$
- $4w_2 - 4 = 0 \implies w_2 = 1$
- $2w_3 - 2w_4 = 0 \implies w_3 = w_4$

The function is minimized at any point of the form $(-1, 1, x, x)$.

(d) No, there is not a unique solution.

11. We are interested in analyzing

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 + \lambda \|w\|^2.$$

(a) To compute $\nabla L(w)$, we compute partial derivatives.

$$\frac{\partial L}{\partial w_j} = \left(\sum_{i=1}^{n} -2x_j^{(i)} (y^{(i)} - w \cdot x^{(i)}) \right) + 2\lambda w_j$$

Thus

$$\nabla L(w) = -2 \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)}) x^{(i)} + 2\lambda w.$$

(b) The update for gradient descent with step size η looks like

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla L(w_t) \\ &= w_t(1 - 2\eta\lambda) + 2\eta \sum_{i=1}^{n} (y^{(i)} - w_t \cdot x^{(i)}) x^{(i)} \end{aligned}$$

(c) The update for stochastic gradient descent looks like the following.

$$w_{t+1} = w_t(1 - 2\eta\lambda) + 2\eta(y^{(i_t)} - w_t \cdot x^{(i_t)}) x^{(i_t)}$$

where i_t is the index chosen at time t.

CSE 250B: Machine learning **Winter 2019**

Homework 4 — Convexity, Perceptron, SVM

1. Are the following functions $f : \mathbb{R} \to \mathbb{R}$ convex, concave, or neither? Justify your answer.
   (a) $f(x) = e^{ax}$, for some constant a.
   (b) $f(x) = |x|$.
   (c) $f(x) = \ln x$, where $x > 0$.
   (d) $f(x) = x^a$, for $a \geq 1$. What if $a \leq 0$? What if $0 \leq a \leq 1$?

2. Show that the following functions $f : \mathbb{R}^d \to \mathbb{R}$ are convex.
   (a) $f(x) = x^T M x$, where $M \in \mathbb{R}^{d \times d}$ is symmetric positive semidefinite.
   (b) $f(x) = e^{u \cdot x}$, for some $u \in \mathbb{R}^d$.
   (c) $f(x) = \max(f_1(x), \ldots, f_k(x))$, where $f_1, \ldots, f_k$ are convex.

3. Recall that the *entropy* of a discrete distribution $p = (p_1, \ldots, p_k)$ over k outcomes is defined as follows:

$$H(p) \;=\; \sum_{i=1}^{n} p_i \log \frac{1}{p_i}.$$

   Show that $H(p)$ is a concave function of p. You may switch to the natural logarithm if you wish.

4. Recall the loss function for regularized least squares: for some constant $\lambda > 0$,

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 + \lambda \|w\|^2,$$

   (a) Obtain an expression for the Hessian $H(w)$: that is, the $d \times d$ matrix of second derivatives.
   (b) Establish that $L(w)$ is a convex function of w.

5. In class, we studied convex *functions*. In this problem, we will define the notion of a convex *set*. Pick any $K \subseteq \mathbb{R}^d$. We say K is a convex set if for any $x, y \in K$, the line segment joining x and y lies entirely in K; more formally, for any $x, y \in K$ and any $0 < \theta < 1$, we have $\theta x + (1 - \theta) y \in K$.

   Which of the following is a convex set?
   (a) The circle: $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.
   (b) The unit ball: $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$.
   (c) A hyperplane: $\{x \in \mathbb{R}^d : w \cdot x = 0\}$ for some $w \in \mathbb{R}^d$.
   (d) All k-sparse points: $\{x \in \mathbb{R}^d : x$ has at most k nonzero coordinates$\}$.
   (e) The set of all $d \times d$ symmetric positive semidefinite matrices (treat each matrix as a vector in $\mathbb{R}^{d(d+1)/2}$).

6. *Norms.* In class, we talked about ℓ_p norms on $\mathbb{R}^d$, which include the following:

   - The l_1 norm: $\|x\|_1 = \sum_{i=1}^d |x_i|$.
   - The l_2 (Euclidean) norm: $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$.
   - The l_∞ norm: $\|x\|_\infty = \max_i |x_i|$.

   We now define norms more generally. A function $f : \mathbb{R}^d \to \mathbb{R}$ is a *norm* if:

   - It is nonnegative: $f(x) \geq 0$ always.
   - $f(x) = 0$ if and only if $x = 0$.
   - It is homogeneous: $f(tx) = |t|f(x)$ for any $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$.
   - It satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$.

   (a) Prove that the ℓ_1 norm satisfies these properties.

   (b) Prove that any norm $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function. (This means we can easily incorporate norms into objective functions we are optimizing.)

   (c) Prove the following two properties. For the second, you may need to use the Cauchy-Schwarz inequality (that is, $|a \cdot b| \leq \|a\|\|b\|$ for any vectors a, b).

   - $\|x\|_1 \geq \|x\| \geq \|x\|_\infty$.
   - $\|x\|_1 \leq \|x\| \cdot \sqrt{d} \leq \|x\|_\infty \cdot d$.

   (d) Another norm that is quite common in machine learning and statistics is the Mahalanobis norm:

   $$\|x\|_A = \sqrt{x^T A x},$$

   where A is a symmetric positive definite matrix. What does the unit ball of this norm, that is $\{x : \|x\|_A \leq 1\}$, look like? *Hint*: think back to the multivariate Gaussian.

7. *A lower bound for the perceptron.* Give an example of a data set $\{(x^{(i)}, y^{(i)})\}$ for which the bound of the perceptron convergence theorem is tight. For convenience, choose the $x^{(i)}$ to have unit length, and show that the number of updates is $1/\gamma^2$.

8. *Small SVM example.* Consider the following small data set in $\mathbb{R}^2$:

   - Points $(1, 2), (2, 1), (2, 3), (3, 2)$ have label -1.
   - Points $(4, 5), (5, 4), (5, 6), (6, 5)$ have label $+1$.

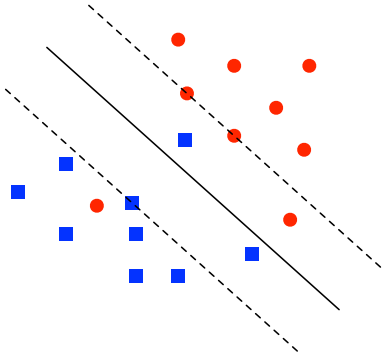
   Now, suppose (hard) SVM is run on this data.

   (a) Sketch the resulting decision boundary.

   (b) What is the (numerical value of the) margin, exactly?

   (c) What are w and b, exactly?

9. *Support vectors.* The picture below shows the decision boundary obtained upon running soft-margin SVM on a small data set of blue squares and red circles.



(a) Mark the support vectors. For each, indicate the approximate value of the corresponding slack variable.

(b) Suppose the factor C in the soft-margin SVM optimization problem were increased. Would you expect the margin to increase or decrease?

CSE 250B: Homework 4 Solutions

1. *Checking convexity/concavity.*

   (a) $f(x) = e^{ax}$ is convex.
   Proof: The second partial derivative $H(x) = f''(x) = a^2 e^{ax} \geq 0$

   (b) $f(x) = |x|$ is convex.
   Proof: $\forall a, b \in \mathbb{R}$ and $\theta \in (0, 1)$,

   $$f(\theta a + (1-\theta)b) = |\theta a + (1-\theta)b| \leq |\theta a| + |(1-\theta)b| = \theta|a| + (1-\theta)|b| = \theta f(a) + (1-\theta)f(b)$$

   (c) $f(x) = \ln x$ is concave.
   Proof: $-f(x) = -\ln x$ is convex because the second derivative

   $$H(x) = -f''(x) = \frac{1}{x^2} \geq 0$$

   (d) $f(x) = x^a \quad (x > 0)$. Here we only consider $x > 0$ because $f(x)$ doesn't always have definition when x is negative. $f(x)$ is convex when $a \geq 1$ and $a \leq 0$, and is concave when $0 < a < 1$.
   Proof: The second derivative

   $$H(x) = a(a-1)x^{a-2}$$

   When $0 < a < 1$, $H(x) < 0$, which means the second derivative of $-f(x)$ is positive, so in this case $f(x)$ is concave. When $a \geq 1$ or $a \leq 0$, $H(x) \geq 0$, so in this case $f(x)$ is convex.

2. *Showing convexity.*

   (a) The Hessian of $f(x) = x^T M x$ is $H(x) = 2M$. Since M is positive semidefinite, so is $2M$; so f is convex.

   (b) The Hessian of $f(x) = e^{u \cdot x}$ is

   $$H(x) = e^{u \cdot x} u u^T,$$

   which can also be written as vv^T, where $v = (e^{u \cdot x}/2)u$. Thus $H(x)$ is P.S.D. and so $f(x)$ is convex.

   (c) Since $f(x) = \max(f_1(x), \ldots, f_k(x))$, where the individual f_i are all convex, we have that for all $x_1, x_2 \in \mathbb{R}$ and $t \in (0, 1)$,

   $$\begin{aligned}
   & f(tx_1 + (1-t)x_2) \\
   = \;& \max\left(f_1(tx_1 + (1-t)x_2), f_2(tx_1 + (1-t)x_2), \ldots, f_k(tx_1 + (1-t)x_2)\right) \\
   \leq \;& \max\left(tf_1(x_1) + (1-t)f_1(x_2), tf_2(x_1) + (1-t)f_2(x_2), \ldots, tf_k(x_1) + (1-t)f_k(x_2)\right) \\
   \leq \;& t\max\left(f_1(x_1), f_2(x_1), \ldots, f_k(x_1)\right) + (1-t)\max\left(f_1(x_2), f_2(x_2), \ldots, f_k(x_2)\right) \\
   = \;& tf(x_1) + (1-t)f(x_2)
   \end{aligned}$$

   Therefore, $f(x)$ is convex.

3. *Entropy.* The negation of the entropy, $N(p) = -H(p)$, has Hessian with entries

   $$\frac{\partial N}{\partial p_i \partial p_j} = \begin{cases} 0 & \text{if } i \neq j, \\ \frac{1}{p_i \ln 2} & \text{if } i = j \end{cases}$$

   This is a diagonal matrix with positive values on the diagonal. Thus the Hessian is P.S.D., whereupon N is convex and H is concave.

4. *Regression problem.*

   (a) Let

$$X = \begin{pmatrix} \leftarrow & x^{(1)} & \rightarrow \\ \leftarrow & x^{(2)} & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & x^{(n)} & \rightarrow \end{pmatrix}$$

   Then we can write the Hessian as

$$H(w) = 2\sum_{i=1}^{n} x^{(i)} \left(x^{(i)}\right)^T + 2\lambda I = 2X^T X + 2\lambda I$$

   (b) For all $z \in \mathbb{R}^d$

$$z^T H z = z^T(2X^T X + 2\lambda I)z = 2(z^T X^T X z + \lambda z^T I z) = 2||Xz||^2 + 2\lambda||z||^2 \geq 0$$

   Therefore, $H(w)$ is P.S.D, which means $L(w)$ is convex.

5. *Convex sets.*

   (a) The circle is not a convex set: for any two points on the circle, the line joining them does not lie on the circle.

   (b) The ball is convex.

   (c) Hyperplanes are convex.

   (d) k-sparse points are not convex: lines joining two such points can be upto $(2k)$-sparse.

   (e) The set of positive semidefinite matrices is closed under addition and multiplication by positive scalars; therefore it is convex.

6. *Norms.*

   (a) We can check that ℓ_1 is a norm by going through the definition, one property at a time:

      i. $\|x\|_1 = \sum_{i=1}^d |x_i| \geq 0$.

      ii. If $x = 0$, then $\|x\|_1 = 0$. If $\exists i, x_i \neq 0$, then $\|x\|_1 \geq |x_i| > 0$. Therefore, $\|x\|_1 = 0$ if and only if $x = 0$.

      iii. For any real-valued t, we have $\|tx\|_1 = \sum_{i=1}^d |tx_i| = |t| \sum_{i=1}^d |x_i| = |t| \, \|x\|_1$

      iv. $\|x + y\|_1 = \sum_{i=1}^d |x_i + y_i| \leq \sum_{i=1}^d |x_i| + |y_i| = \sum_{i=1}^d |x_i| + \sum_{i=1}^d |y_i| = \|x\|_1 + \|y\|_1$

   (b) Invoking homogeneity and the triangle inequality, we have that for any norm f,

$$f(\theta x + (1-\theta)y) \leq f(\theta x) + f((1-\theta)y) = |\theta| f(x) + |1-\theta| f(y) = \theta f(x) + (1-\theta) f(y).$$

   Thus any norm is a convex function.

   (c) Various inequalities relating $\|x\|_1$, $\|x\|$, and $\|x\|_\infty$:

      i. $\|x\|_1 = \sqrt{(\sum_{i=1}^d |x_i|)^2} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d |x_i||x_j|} \geq \sqrt{\sum_{i=1}^d x_i^2} = \|x\|$.
      $\|x\| = \sqrt{\sum_{i=1}^d x_i^2} \geq \sqrt{\max_i x_i^2} = \max_i |x_i| = \|x\|_\infty$

      ii. Let vector $a = (|x_1|, |x_2|, \dots, |x_d|), b = (1, 1, \dots, 1)_d$
      $\|x\|_1 = \sum_{i=1}^d |x_i| = |a \cdot b| \leq \|a\| \, \|b\| = \sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d 1^2} = \|x\| \cdot \sqrt{d}$.
      $\|x\| = \sqrt{\sum_{i=1}^d x_i^2} \leq \sqrt{d \cdot \max_i x_i^2} = \|x\|_\infty \cdot \sqrt{d}$.
      Therefore, $\|x\|_1 \leq \|x\| \cdot \sqrt{d} \leq \|x\|_\infty \cdot d$.

(d) The unit ball $\{x : x^T Ax \leq 1\}$ is an ellipsoid.

7. *A lower bound for the perceptron.* Pick any $\gamma > 0$. Consider the following data set in $\mathbb{R}^d$, where $d = 1/\gamma^2$:

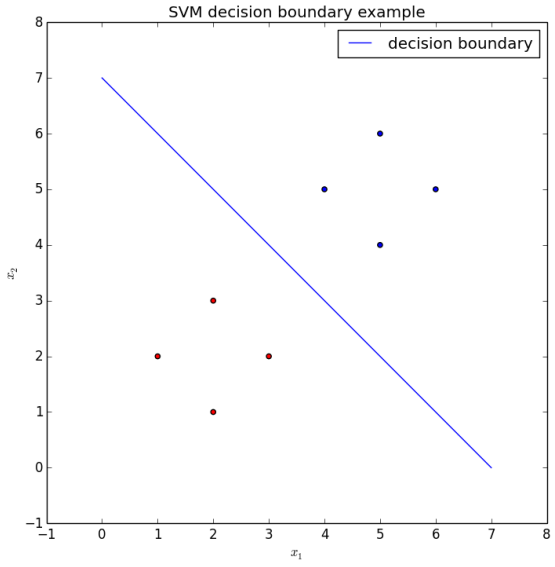
   - There are d points, each corresponding to one coordinate direction: $e_1, e_2, \ldots, e_d$, where e_i is the vector with all zeros except for a 1 at position i.
   - All points have label $+1$.

   These points are correctly classified by the vector $w^* = (\gamma, \gamma, \ldots, \gamma)$, which has unit length and has margin $\min_i(w^* \cdot e_i) = \gamma$.

   Now suppose the perceptron algorithm is run on this data set, and that it produces a linear separator w. If perceptron does not update on e_i, then $w_i = 0$ and w will not correctly classify e_i. Therefore, there must be at least one update for every data point: a total of $1/\gamma^2$ updates.

8. *Small SVM example.*

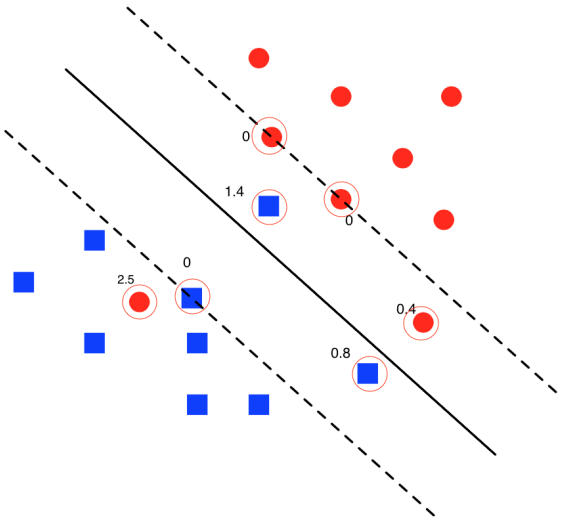
   (a)



   (b) The margin is $\sqrt{2}$.

   (c) w lies in the direction $\begin{pmatrix}1\\1\end{pmatrix}$ and has length $1/\sqrt{2}$ (since the margin is $\sqrt{2}$); therefore, $w = \begin{pmatrix}1/2\\1/2\end{pmatrix}$. We know that the point $x_o = (4, 3)$ lies on the decision boundary; solving $w \cdot x_o + b = 0$ yields $b = -7/2$.

9. *Support vectors.* The margin decreases if the factor C is increased.


0
1.4
0
0
2.5
0.4
0.8

CSE 250B: Machine learning **Winter 2019**

Homework 5 — Kernels and Decision Trees

Problem 1

In the following problems, suppose that K, K_1 and K_2 are kernels with feature maps ϕ, ϕ^1 and ϕ^2. For the following functions $K'(x, z)$, state if they are kernels or not. If they are kernels, write down the corresponding feature map, in terms of ϕ, ϕ^1, ϕ^2 and c, c_1, c_2. If they are not kernels, prove that they are not.

1. $K'(x, z) = cK(x, z)$, for $c > 0$. True.
2. $K'(x, z) = cK(x, z)$, where $c < 0$, and there exists some x for which $K(x, x) > 0$. False.
3. $K'(x, z) = c_1K_1(x, z) + c_2K_2(x, z)$ for $c_1, c_2 > 0$. True.
4. $K'(x, z) = K_1(x, z)K_2(x, z)$. True. Product of two kernels is a kernel.

Problem 2

For the following functions $K(x, z)$, state if it is a kernel or not. If the function is a kernel, then write down its feature map. If it is not a kernel, prove that it is not one. For your proof, you can use the answers to Problem 1.

1. $x = [x_1, x_2]$, $z = [z_1, z_2]$, x_1, x_2, z_1, z_2 are real numbers. $K(x, z) = x_1z_2$. False. x = [1,-1] => K(x,x)<0
2. Let $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = 1 - \langle x, z\rangle$. False. x = [1,0,0] z = [2,0,0,0]
3. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = \|x - z\|^2$. False. Cauchy-schwarz inequality
4. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, and f is a function. $K(x, z) = f(x_1, x_2)f(z_1, z_2)$. True
5. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x, z) = \frac{1-\langle x,z\rangle^2}{1-\langle x,z\rangle}$. True
6. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \min(x_i, z_i)$. True
7. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. True

$$K(x, z) = (1 + x_1z_1)(1 + x_2z_2) \ldots (1 + x_dz_d)$$

8. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \max(x_i, z_i)$.

False. Contradictory with Cauchy-Swartz inequality.

Problem 3

A group of biologists would like to determine which genes are associated with a certain form of liver cancer. After much research, they have narrowed the possibilities down to two genes, let us call them A and B. After analyzing a lot of data, they have also calculated the following joint probabilities.

	Cancer	No Cancer
Gene A	$\frac{1}{2}$	$\frac{1}{10}$
No Gene A	$\frac{1}{5}$	$\frac{1}{5}$

	Cancer	No Cancer
Gene B	$\frac{2}{5}$	$\frac{3}{20}$
No Gene B	$\frac{3}{10}$	$\frac{3}{20}$

1. Let X denote the 0/1 random variable which is 1 when a patient has cancer and 0 otherwise. Let Y denote the 0/1 random variable which is 1 when gene A is present, 0 otherwise, and let Z denote the 0/1 random variable which is 1 when gene B is present and 0 otherwise. Write down the conditional distributions of $X|Y = y$ for $y = 0, 1$ and $X|Z = z$, for $z = 0, 1$.
2. Calculate the conditional entropies $H(X|Y)$ and $H(X|Z)$.
3. Based on these calculations, which of these genes is more informative about cancer?

Problem 4

Since a decision tree is a classifier, it can be thought of as a function that maps a feature vector x in some set $\mathcal{X}$ to a label y in some set $\mathcal{Y}$. We say two decision trees T and T' are *equal* if for all $x \in \mathcal{X}$, $T(x) = T'(x)$.

The following are some statements about decision trees. For these statements, assume that $\mathcal{X} = \mathbb{R}^d$, that is, the set of all d-dimensional feature vectors. Also assume that $\mathcal{Y} = \{1, 2, \ldots, k\}$. Write down if each of these statements are correct or not. If they are correct, provide a brief justification or proof; if they are incorrect, provide a counterexample to illustrate a case when they are incorrect.

1. If the decision trees T and T' do not have exactly the same structure, then they can never be equal.
2. If T and T' are any two decision trees that produce zero error on the same training set, then they are equal.

Problem 5

In this problem, we will formally examine how transforming the training data in simple ways can affect the performance of common classifiers. Transforming training features by scaling is equivalent to measuring these features in different units; in practice, we frequently have to combine multiple homogeneous or heterogeneous features, and it is important to understand how changing units in which features are measured can affect machine learning algorithms.

Suppose we are given a training data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where each feature vector x_i lies in d-dimensional space. Suppose each $x_i = [x_i^1, x_i^2, \ldots, x_i^d]$, so coordinate j of x_i is denoted by x_i^j.

For each x_i, suppose we transform it to z_i by rescaling each axis of the data by a fixed factor; that is, for every $i = 1, \ldots, n$ and every coordinate $j = 1, \ldots, d$, we write:

$$z_i^j = \alpha^j x_i^j$$

Here α^js are real, non-zero and positive constants. Thus, our original training set S is transformed after rescaling to a new training set $S' = \{(z_1, y_1), \ldots, (z_n, y_n)\}$. For example, if we have two features, and if $\alpha^1 = 3$, and $\alpha^2 = 2$, then, a feature vector $x = (x^1, x^2)$ gets transformed by rescaling to $z = (z^1, z^2) = (3x^1, 2x^2)$.

A classifier $C(x)$ in the original space (of x's) is said to be equal to a classifier $C'(z)$ in the rescaled space (of z's) if for every $x \in \mathbb{R}^d$, $C(x) = C'(z)$, where z is obtained by transforming x by recaling. In our previous example, the classifier C in the original space:

$$C(x): \textit{ Predict } 0 \textit{ if } x^1 \leq 1, \textit{ else predict } 1.$$

is equal to the classifier C' in the rescaled space:

$$C'(z): \textit{ Predict } 0 \textit{ if } z^1 \leq 3, \textit{ else predict } 1.$$

This is because if $C(x) = 0$ for an $x = (x^1, x^2)$, then $x^1 \leq 1$. This means that for the transformed vector $z = (z^1, z^2) = (3x^1, 2x^2)$, $z^1 = 3x^1 \leq 3$, and thus $C'(z) = 0$ as well. Similarly, if $C(x) = 1$, then $x^1 > 1$ and $z^1 > 3$ and thus $C(z) = 1$. Now, answer the following questions:

1. First, suppose that all the α^i values are equal; that is, $\alpha^1 = \ldots = \alpha^d$. Suppose we train a k-NN classifier C on S and a k-NN classifier C' on S'. Are these two classifiers equal? What if we trained C and C' on S and S' respectively using the ID3 Decision Tree algorithm? What if we trained C and C' on S and S' respectively using the Perceptron algorithm? If the classifiers are equal, provide a *brief* argument to justify why; if they are not equal, provide a counterexample.

2. Repeat your answers to the questions in part (1) when the α_is are different. Provide a *brief* justification for each answer if the classifiers are equal, and a counterexample if they are not.

3. From the results of parts (1) and (2), what can you conclude about how k-NN, decision trees and perceptrons behave under scaling transformations?

CSE 250B: Homework 5 Solutions

Problem 1

1. Suppose $K(x,z) = \langle \phi(x), \phi(z) \rangle$ for some feature map ϕ, and let $\phi'(x) = \sqrt{c}\phi(x)$. Then, for all x and z,

$$K'(x,z) = cK(x,z) = c\langle \phi(x), \phi(z) \rangle = \langle \sqrt{c}\phi(x), \sqrt{c}\phi(z) \rangle$$

   Therefore $K'(x,z)$ is a kernel corresponding to the feature map ϕ'.

2. Suppose x_0 is the x for which $K(x,x) > 0$. Consider the 1×1 kernel matrix $K' = K'(x_0, x_0)$ for the kernel K' and the data point x_0. Then, $K' = cK(x_0, x_0)$. If $z = 1$, then $z^\top K' z = cK(x_0, x_0) < 0$, which violates the kernel Positive Semi Definiteness (PSD) property. Thus K' is not a kernel.

3. Suppose $K_1(x,z) = \langle \phi^1(x), \phi^1(z) \rangle$ and $K_2(x,z) = \langle \phi^2(x), \phi^2(z) \rangle$. Then, for all x and z,

$$\begin{aligned} K'(x,z) &= c_1\langle \phi^1(x), \phi^1(z) \rangle + c_2\langle \phi^2(x), \phi^2(z) \rangle = \langle \sqrt{c_1}\phi^1(x), \sqrt{c_1}\phi^1(z) \rangle + \langle \sqrt{c_2}\phi^2(x), \sqrt{c_2}\phi^2(z) \rangle \\ &= \langle \phi'(x), \phi'(z) \rangle \end{aligned}$$

   where $\phi'(x)$ is a concatenation of the feature maps $\sqrt{c_1}\phi^1(x)$ and $\sqrt{c_2}\phi^2(x)$. In other words, if the feature maps ϕ^1 and ϕ^2 have m_1 and m_2 coordinates respectively, then ϕ' has $m_1 + m_2$ coordinates; for any x, the first m_1 coordinates of $\phi'(x)$ are $\sqrt{c_1}\phi^1_1(x), \sqrt{c_1}\phi^1_2(x), \ldots, \sqrt{c_1}\phi^1_{m_1}(x)$ and the remaining m_2 coordinates of $\phi'(x)$ are $\sqrt{c_2}\phi^2_1(x), \sqrt{c_2}\phi^2_2(x), \ldots, \sqrt{c_2}\phi^2_{m_2}(x)$. Therefore $K'(x,z)$ is a kernel corresponding to the feature map ϕ'.

4. Suppose $K_1(x,z) = \langle \phi^1(x), \phi^1(z) \rangle$ and $K_2(x,z) = \langle \phi^2(x), \phi^2(z) \rangle$. If x and z are d-dimensional vectors, then, for all x and z,

$$\begin{aligned} K'(x,z) &= K_1(x,z)K_2(x,z) = \langle \phi^1(x), \phi^1(z) \rangle \cdot \langle \phi^2(x), \phi^2(z) \rangle \\ &= \left(\sum_i \phi^1_i(x)\phi^1_i(z) \right) \cdot \left(\sum_j \phi^2_j(x)\phi^2_j(z) \right) = \sum_{i,j=1}^{d} (\phi^1_i(x)\phi^2_j(x)) \cdot (\phi^1_i(z)\phi^2_j(z)) \\ &= \langle \phi'(x), \phi'(z) \rangle \end{aligned}$$

   where

$$\phi'(x) = \begin{bmatrix} \phi^1_1(x)\phi^2_1(x) \\ \phi^1_1(x)\phi^2_2(x) \\ \phi^1_2(x)\phi^2_1(x) \\ \phi^1_1(x)\phi^2_3(x) \\ \phi^1_2(x)\phi^2_2(x) \\ \phi^1_3(x)\phi^2_1(x) \\ \vdots \end{bmatrix} \tag{1}$$

   That is, ϕ' is a $d^2 \times 1$ feature map, which has a coordinate $\phi_{(i,j)}(\cdot)$ corresponding to each pair (i,j), $1 \le i, j \le d$, where $\phi_{(i,j)}(x) = \phi^1_i(x)\phi^2_j(x)$. Thus $K'(x,z)$ is a kernel corresponding to the feature map ϕ'.

Problem 2

1. $K(x,z)$ is not a kernel.
   For $x = [1, -1]$, we have $K(x,x) = 1 \times -1 = -1$. The corresponding kernel matrix $K = -1$. For $v = 1$, $v^\top K v = -1 < 0$, which violates the PSD property. Thus K is not a kernel.

2. $K(x, z)$ is not a kernel.
For $x = [2, 2, \cdots]$, we have $K(x, x) = 1 - \langle x, x \rangle = 1 - 4d$. The corresponding kernel matrix $K = 1 - 4d$.
For $v = 1$, $v^\top K v = 1 - 4d < 0$, which violates the kernel PSD property for $d > 0$. Thus K is not a kernel.

3. $K(x, z)$ is not a kernel.
One way to prove that K is not a kernel is to show a counterexample to the PSD property. Pick $x = [1, 0, \ldots, 0], z = [2, 0, \ldots, 0], v = [1, -1]^\top$. Then the kernel matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and $v^\top A v = -2 < 0$, which violates positivity.

A nice, second way to prove this is through contradiction. Suppose K a kernel, such that $K(x, z) = \langle \phi(x), \phi(z) \rangle$. Recall the Cauchy-Schwarz Inequality for inner product, that we discussed in Lecture 2:

$$\langle \phi(x), \phi(z) \rangle^2 \leq \langle \phi(x), \phi(x) \rangle \cdot \langle \phi(z), \phi(z) \rangle \tag{2}$$

From this inequality,

$$K(x, z)^2 \leq K(x, x) \cdot K(z, z) \tag{3}$$

Suppose x is any vector with norm 1 and let $z = 2x$. By the definition of K, we have $K(x, x) = \|x - x\|^2 = 0$ and $K(x, x) \cdot K(z, z) = 0$. However $K(x, z) = \|x - 2x\|^2 = 1 > K(x, x) \cdot K(z, z)$, which leads to a contradiction! Thus K is not a kernel.

4. $K(x, z)$ is a kernel corresponding to the feature map $\phi(x) = f(x_1, x_2)$.

5. $K(x, z)$ is a kernel.
Recall that $a^2 - b^2 = (a - b) \cdot (a + b)$

Hence, we have

$$\frac{1 - \langle x, z \rangle^2}{1 - \langle x, z \rangle} = 1 + \langle x, z \rangle \tag{4}$$

In the above equation, we can rewrite 1 as $\langle x, z \rangle^0$

Thus, we can now write, $K(x, z) = K_0(x, z) + K_1(x, z)$. In Problem 1, we saw that the sum or product of two kernels is also a kernel. We know that $K_0(x, z)$ and $K_1(x, z)$ are both kernels.

The feature map $\phi_0(x)$ corresponding $K_0(x, z)$ is

$$\phi_0(x) = 1 \tag{5}$$

$K_1(x, z)$ corresponds to the feature map

$$\phi_1(x) = x \tag{6}$$

Using Problem 1 Part 3, $K(x, z) = K_0(x, z) + K_1(x, z)$ is a kernel corresponding to the feature map ϕ', where for any x, $\phi'(x)$ is a concatenation of the feature maps $\phi_0(x)$ and $\phi_1(x)$.

6. $K(x, z)$ is a kernel.
Let $K_i(x, z) = \min(x_i, z_i)$. From Problem 1, we know that the sum of two kernels K_1 and K_2 is also a kernel whose corresponding feature map is the concatenation of the feature maps corresponding to K_1 and K_2. Thus if we can find the feature maps for all $K_i(x, z)$, then we can get the feature map for $K(x, z)$ by concatenating these maps. Consider following feature map:

$$\phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top \tag{7}$$

where $f_k(t) = I(t \geq k) = \begin{cases} 1 & t \geq k \\ 0 & t < k \end{cases}$. Without loss of generality, suppose that $x_i \leq z_i$. Then $\phi_i(x) = [1, \ldots, 1, 0, \ldots, 0]^\top$ where only the first x_i entries are 1. Analogously, $\phi_i(z) = [1, \ldots, 1, 0, \ldots, 0]^\top$ where only the first z_i entries are 1. Then

$$\langle \phi_i(x), \phi_i(z) \rangle = \sum_{i=1}^{x_i} 1 \cdot 1 + \sum_{i=x_i+1}^{z_i} 0 \cdot 1 + \sum_{i=z_i+1}^{100} 0 \cdot 0 = x_i = \min(x_i, z_i)$$

Therefore $K_i(x, z)$ is a kernel corresponding to the feature map $\phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top$, and $K(x, z)$ is a kernel corresponding to the feature map $\phi(x)$ which is a concatenation of the feature maps $\phi_1(x), \phi_2(x), \ldots, \phi_d(x)$.

7. $K(x, z)$ is a kernel.

   Let $K_i(x, z) = 1 + x_i z_i$, then $K(x, z) = \prod_{i=0}^{d} K_i(x)$. From Problem 1, we know that the product of two kernels is also a kernel. Since $K_i(x, z)$ is a kernel corresponding to the feature map $\phi_i(x) = [1, x_i]^\top$, $K(x, z)$ is also a kernel. More specifically, $K(x, z)$ is a kernel corresponding to the feature map $\phi(x)$, where for any x, $\phi(x)$ has 2^d coordinates, one corresponding to each subset S of $\{1, 2, \ldots, d\}$. $\phi_S(x)$, the coordinate of $\phi(x)$ corresponding to the set S is $\prod_{i \in S} x_i$. This kernel is called the *All Subsets* kernel.

8. $K(x, z)$ is not a kernel.
   One way to prove this is by showing a violation of the PSD property. Let $x = [0, \ldots, 0], z = [1, 0, \ldots, 0]$ and $v = [1, -1]^\top$. Then the kernel matrix

   $$K = \begin{bmatrix} K(x, x) & K(x, z) \\ K(z, x) & K(z, z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

   Thus, $v^\top A v = -1 < 0$, which violates positivity.

   Another nice way is through a violation of the Cauchy-Schwartz inequality. Consider $x = [0, \ldots, 0]$ and $z = [1, 0, \ldots, 0]$. Then $K(x, x) = 0$, $K(x, z) = K(z, z) = 1$, which violates Cauchy-Schwarz inequality – that is $K(x, z)^2 \geq K(x, x) \cdot K(z, z)$.

Problem 3

1. First, we can compute the marginal distributions of Y and Z as follows,

y	0	1
$P(Y = y)$	$\frac{2}{5}$	$\frac{3}{5}$

z	0	1
$P(Z = z)$	$\frac{9}{20}$	$\frac{11}{20}$

   Then, by definition of conditional probability, i.e. $P(X = x|Y = y) = \dfrac{P(X = x, Y = y)}{P(Y = y)}$, we can get the conditional distributions of $X|Y$ as follows.

x	0	1
$P(X = x\|Y = 0)$	$\frac{1}{2}$	$\frac{1}{2}$
$P(X = x\|Y = 1)$	$\frac{1}{6}$	$\frac{5}{6}$

   Similarly we have the conditional distributions of $X|Z$ as follows,

x	0	1
$P(X = x\|Z = 0)$	$\frac{1}{3}$	$\frac{2}{3}$
$P(X = x\|Z = 1)$	$\frac{3}{11}$	$\frac{8}{11}$

2. By the definition of conditional entropy, $H(X|Y) = P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1)$.

$$\begin{aligned} H(X|Y=0) &= -P(X=0|Y=0)\log P(X=0|Y=0) - P(X=1|Y=0)\log P(X=1|Y=0) \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \\ &= \log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Y=1) &= -P(X=0|Y=1)\log P(X=0|Y=1) - P(X=1|Y=1)\log P(X=1|Y=1) \\ &= -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} \\ &= \log 6 - \frac{5}{6}\log 5 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Y) &= P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\left(\log 6 - \frac{5}{6}\log 5\right) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\log 6 - \frac{1}{2}\log 5 \end{aligned}$$

For $H(X|Z)$, we can get

$$\begin{aligned} H(X|Z=0) &= -P(X=0|Z=0)\log P(X=0|Z=0) - P(X=1|Z=0)\log P(X=1|Z=0) \\ &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= \log 3 - \frac{2}{3}\log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Z=1) &= -P(X=0|Z=1)\log P(X=0|Z=1) - P(X=1|Z=1)\log P(X=1|Z=1) \\ &= -\frac{3}{11}\log\frac{3}{11} - \frac{8}{11}\log\frac{8}{11} \\ &= \log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Z) &= P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) \\ &= \frac{9}{20}\left(\log 3 - \frac{2}{3}\log 2\right) + \frac{11}{20}\left(\log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8\right) \\ &= -\frac{3}{2}\log 2 + \frac{3}{10}\log 3 + \frac{11}{20}\log 11 \end{aligned}$$

Using natural logarithm, the numerical values are shown as follows.

$H(X\|Y=0)$	0.693147180560
$H(X\|Y=1)$	0.450561208866
$H(X\|Y)$	0.547595597544
$H(X\|Z=0)$	0.63651416829
$H(X\|Z=1)$	0.5859526183
$H(X\|Z)$	0.6087053158

3. From the table above, $H(X|Y) < H(X|Z)$. This suggests that there is less uncertainty in X when given Y than when given Z. Therefore gene A is more informative about the cancer.

Problem 4

1. False.
   Counterexample: Consider a classifier for data which uses one feature (called Feature1).

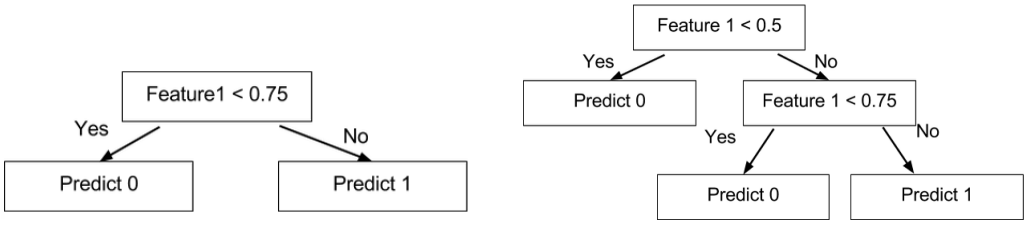


Figure 1: Two Decision Trees which are equal (see definition in question) but have different structures

2. False.
   If T and T' produce zero error on the same training set $S \subseteq \mathcal{X}$, then, $\forall x \in S$, $T(x) = T'(x)$. However, the training set typically does not include all elements in feature space $\mathcal{X}$. Thus, there exist such $x_0 \in \mathcal{X} - S$ that $T(x_0) \neq T'(x_0)$. For example, consider the following training set:

Feature 1	Feature 2	Label
0	0	0
1	1	1

   For training set above, the two decision trees shown in Figure 2 both produce zero error. However, for the point $x_1 = (0, 1)$ or the point $x_2 = (1, 0)$, these two trees would give different predictions. Hence they are not equal.

Figure 2: Two Decision Trees with Zero Error on S

Problem 5

For this question, we assume that ties are always broken in a consistent manner for both the k-NN and ID3 decision tree algorithms.

k-NN. We will obtain equal k-NN classifiers before and after a space transform for an arbitrary data set, if and only if, the following *relative distance* conditon holds: for any three points x, x_p and x_q in the original space, $d(x, x_p) \geq d(x, x_q)$ implies $d(z, z_p) \geq d(z, z_q)$, whre z, z_p and z_q are the points after rescaling. In other words, we need to ensure that in all cases, the nearest neighbors of a point in the original space are still the nearest neighbors in the rescaled space.

In the case of a uniform scaling factor (all $\alpha^j = \alpha$), the distance between any two points z_1 and z_2 in the rescaled space is,

$$d(z_1, z_2) = \sqrt{\sum_j (z_1^j - z_2^j)^2} = \sqrt{\sum_j (\alpha x_1^j - \alpha x_2^j)^2} = \alpha \sqrt{\sum_j (x_1^j - x_2^j)^2} = \alpha d(x_1, x_2),$$

This is simply the distance in the original space scaled by a constant α. Clearly the relative distance condition holds. In particular, this means that the training points that are the nearest neighbors of x in the original space remain the nearest neighbors of z in the rescaled space, therefore prediction for x remains the same as the prediction for z.

For nonuniform scaling factors, the relative distance condition does not necessarily hold. One extreme example is if $\alpha^1 = 1, \alpha^2 = 0.0001$ (a very small quantity). The transform now essentially projects each point to the x-axis (although the point will not be exactly on the axis). Consider the training points $(1, 0)$ with label 0 and $(0, 1)$ with label 1, and a test example $(0.1, 0)$. In the original space, $(0.1, 0)$ is closer to $(1, 0)$ than $(0, 1)$ and will be assigned label 0; in the rescaled space however, it will be rescaled to be $(0.1, 0)$, will be closer to $(0, 1)$ (now rescaled as $(0, 0.001)$), and thus will be assigned label 1 by the 1-NN classifier. Therefore in this case, we are not guaranteed to get the same k-NN classisifer.

ID3 Decision Tree. The decision trees produced by the ID3 algorithm will be equal in both cases, assuming that ties are broken in a consistent manner. We can show this by induction. In what follows, we will say that a splitting rule (j, t) in the original space *is equal to* to a splitting rule $(j, \alpha^j t)$ in the rescaled space.

We run the ID3 algorithm on S and S' simultaneously, and maintain the following invariants at each step of the algorithm. If T and T' are the trees built based on S and S' respectively, then, (a) T and T' have the same structure, (b) for each internal node v in T, the splitting rule at v is equal to the splitting rule at the corresponding internal node v' in T' and (c) if D is the dataset associated with a leaf node in T, then the dataset associated with the corresponding leaf node in T' is the rescaled version of points in D.

The invariant holds at the beginning of the algorithm, as the only (leaf) node is the root, which is associated with S in T and S' in T'. Suppose the invariant holds at step t of the algorithm, and at step $t + 1$ we split a node v in T such that the dataset associated with v is D. If the splitting rule used is (j, t), then, this splitting rule has the highest information gain among all the possible splitting rules. Observe that as the corresponding node v' in T' is associated with a scaled version D' of D, for any j and t, the information gain of a splitting rule $(j, \alpha^j t)$ at v' is equal to the information gain of the splitting rule (j, t) in v. Thus, assuming that ties are broken consistently, we will pick the splitting rule $(j, \alpha^j t)$ to split node v'. Thus invariants (a) and (b) are maintained after step $t + 1$. Finally, invariant (c) is also maintained as the subset of D for which feature j is $\leq t$ is exactly equal to the subset of D' for which feature j is $\leq \alpha^j t$.

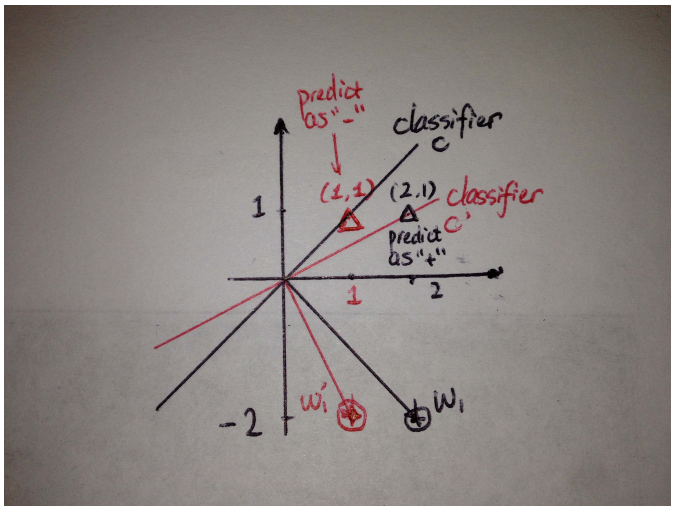
Thus, at the end of the ID3 decision tree algorithm, we arrive at two trees T and T' which have exactly the same structure, where the corresponding nodes v and v' have equal splitting rules. Thus if a test example x follows a path P in T from the root to the leaf, its rescaled version z will follow exactly the same path in T' from root to leaf and will be classified the same way. Therefore the two decision trees will be equal.

Perceptron. For a uniform scaling factor α, we claim that at any step, if the hyperplane normal in the original space is w, then the hyperplane normal in the rescaled space must be αw. If this claim is true, then the classfiers in the two spaces will be equal, because as $\alpha > 0$,
$\text{sign}(\langle w, x_t \rangle) = \text{sign}(\langle \alpha w, z_t \rangle) = \text{sign}(\langle \alpha w, \alpha x_t \rangle)$.

We prove this by induction. The base case is trivial because w is initialized to 0 in both spaces. Then suppose our claim is true for step $t - 1$, we show that the claim still holds at step t. At step t the algorithm predicts the label for the training data (x_t, y_t) in the original space and training data $(z_t = \alpha x_t, y_t)$ in the rescaled space. It is easy to see that the prediction result is the same for the classifiers in both spaces as $\alpha > 0$. If the result is correct, then no change is made to either w. If the result is wrong, the normal in the original space is updated to $w + y_t x_t$, while in the rescaled space, the normal is updated to $\alpha w + y_t z_t = \alpha(w + y_t x_t)$. Thus the claim still holds at this step. Therefore the Perceptron algorithm produces equal classifiers in both spaces.

For non-uniform α's, the two classifiers are not equal. One counter-example is given below. There is only one positive training data $(2, -2)$, which becomes $(1, -2)$ in the rescaled space. Consider the test data $(2, 1)$, and the rescaled version $(1, 1)$. The resulting classifier classifies them into different labels.

Behavior under scaling transformations. In case of uniform scaling transformations (same α^i) across all features/dimensions, all the 3 algorithms are equally robust. However, in case of non-uniform scaling transformations (different α^i), ID3 Decision Trees are more robust to compared to k-NN and Perceptrons.



CSE 250B: Machine learning **Winter 2018**

Homework 6 — More Convex programs and generalization theory

1. We are given a set of m equations in n unknowns $x_1, \ldots, x_n$:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

   It might not be possible to satisfy all these equations exactly; what we want is to find a solution $x = (x_1, \ldots, x_n)$ such that the maximum deviation

$$\max_{1 \le i \le m} \left| b_i - \sum_{j=1}^{n} a_{ij} x_j \right|$$

   is as small as possible. Write this as a linear program.

2. A *halfspace* in $\mathbb{R}^d$ is specified by a vector $w \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$, and is defined as $\{x : w \cdot x \le b\}$.

   (a) Now suppose we have a collection of halfspaces, given by $w_1, w_2, \ldots$ and $b_1, b_2, \ldots$, respectively. There might be infinitely many of them. Show that their intersection is a convex set.

   (b) Can you express the unit ball $\{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ as the intersection of infinitely many halfspaces?

3. We are given two *polyhedra* $P_1, P_2 \subseteq \mathbb{R}^d$, each specified as the intersection of finitely many halfspaces. We would like to find the distance between these two bodies: the smallest possible value $\|x_1 - x_2\|$, where $x_1 \in P_1$ and $x_2 \in P_2$. Show how to express this as a convex program.

4. Let $\mathcal{X} = \{0,1\}^d$. The class $\mathcal{H}$ of *monotone disjunctions* consists of classifiers h that are given by a disjunction (logical OR) of some subset of the d features. For instance, the classifier

$$h(x) = x_1 \vee x_3 \vee x_8$$

   assigns label 1 to points $x \in \mathcal{X}$ for which any of the features x_1, x_3, x_8 are set; and assigns label 0 otherwise. Suppose we obtain a training set of n points, drawn i.i.d. from an unknown underlying distribution, and we find a monotone disjunction $h \in \mathcal{H}$ that is correct on all n points. We would like to give a bound on the true error of h.

   (a) What is $|\mathcal{H}|$? Your answer should be a function of d.

   (b) Give a bound on the true error of h that holds with probability at least $1 - \delta$ over the choice of training data.

   (c) What bound could you give if instead we looked at the smaller class $\mathcal{H}_k \subset \mathcal{H}$ of *k-sparse monotone disjunctions*: that is, monotone disjunctions consisting of at least 1 and at most k variables?

5. *Estimating the bias of a coin.* A coin of bias 3/4 is tossed 300 times and an empirical estimate $\widehat{p}$ of the bias is obtained. Use the central limit theorem to come up with an interval in which $\widehat{p}$ will lie, with 95% probability.

6. Determine the VC dimension of the following concept classes. Justify your answers.

   (a) *Intervals on the line.* $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = 1(a \leq x \leq b)$.

   (b) *Axis-aligned rectangles in the plane.* Each $h \in \mathcal{H}$ is given by an axis-aligned rectangle in $\mathbb{R}^2$, where points inside the rectangle are labeled 1, and points outside are labeled 0.

7. *Isotonic regression.* In a *line fitting* problem, we have a data set consisting of pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ and we want to draw a line through them. More precisely, we want to find parameters $a, b \in \mathbb{R}$ such that $f(x) = ax + b$ passes as close as possible to the points. We have already seen a least-squares formulation of this.

   In *isotonic regression*, we allow a more general function f. It doesn't have to be a line: it just needs to be monotonically increasing, that is, $f(x) \geq f(x')$ whenever $x \geq x'$.

   (a) Here is a training set of six points (x_i, y_i):

   $$(4, 20),\ (2, 5),\ (5, 9),\ (3, 7),\ (1, 10),\ (6, 12).$$

   Plot these points, and sketch a function $f : \mathbb{R} \to \mathbb{R}$ that is monotonically increasing and that passes through *as many of these points as possible.*

   Let's sort the data points so that $x_1 \leq x_2 \leq \cdots \leq x_n$. Monotonicity then means

   $$f(x_1) \leq f(x_2) \leq \cdots \leq f(x_n).$$

   In fact, we can choose any $f(x_i)$ values that meet this requirement; and we can fill in the rest of the f-curve by, say, linearly interpolating between these points.

   How shall we evaluate candidate functions f? In part (a), we used the loss function

   $$L_o(f) \ = \ \#\text{ of training points that } f \text{ does not pass through.}$$

   Finding the optimal such f is called the *longest increasing subsequence* problem in computer science, and can be solved efficiently. However, we typically prefer to use a different, *least-squares* loss.

   Here is a least-squares formulation of our problem: given $x_1 \leq x_2 \leq \cdots \leq x_n$ and corresponding values $y_1, \ldots, y_n$, find $f_1, f_2, \ldots, f_n \in \mathbb{R}$ such that $f_1 \leq f_2 \leq \cdots \leq f_n$ and such that the squared loss

   $$L(f) = \sum_{i=1}^{n} (y_i - f_i)^2$$

   is minimized. (Here f_i corresponds to $f(x_i)$.)

   (b) Show that this is a convex optimization problem.

   An elegant approach to solving this problem is the *pool adjacent violators* algorithm. It starts by simply setting $f_i = y_i$ for all i, and then repeatedly fixes any monotonicity violations: any time it finds $f_i > f_{i+1}$, it resets both of them to the average of f_i, f_{i+1} and merges points x_i and x_{i+1}.

   Here is the algorithm, given a set of x values and their corresponding $y(x)$.

- Let S be the sorted list of x-values
- For all x in S:
  - Set $f(x) = y(x)$
  - Assign weight $w(x) = 1$
- While there adjacent values $x < x'$ in S with $f(x) > f(x')$:
  - Remove x' from S and set a pointer from it to x
  - Let $f(x) = \frac{w(x)f(x)+w(x')f(x')}{w(x)+w(x')}$
  - Let $w(x) = w(x) + w(x')$

At the end, each of the original x-points either lies in the list S, in which case it receives value $f(x)$, or leads to some $\tilde{x}$ in list S by following pointers, in which case it receives value $f(\tilde{x})$.

(c) Run this algorithm on the small data set of six points from part (a). What values of f does it yield?

CSE 250B: Homework 6 Solutions

1. Here is a linear program, over variables $x \in \mathbb{R}^n$ and $v \in \mathbb{R}$:

$$\begin{aligned} &\min v \\ -b_i + \sum_{j=1}^{n} a_{ij} x_j &\leq v, \quad i = 1, 2, \ldots, m \\ b_i - \sum_{j=1}^{n} a_{ij} x_j &\leq v, \quad i = 1, 2, \ldots, m \end{aligned}$$

2. (a) Let K denote the intersection of halfspaces given by $w_1, w_2, \ldots \in \mathbb{R}^d$ and $b_1, b_2, \ldots \in \mathbb{R}$:

$$K \;=\; \bigcap_i \{x : w_i \cdot x \leq b_i\}.$$

For any $x, y \in K$ and $0 < \theta < 1$,

$$w_i \cdot (\theta x + (1-\theta)y) \;=\; \theta w_i \cdot x + (1-\theta) w_i \cdot y \;\leq\; \theta b_i + (1-\theta) b_i \;=\; b_i, \quad \text{for } i = 1, 2, \ldots$$

Therefore, $\theta x + (1-\theta)y \in K$; and K is a convex set.

(b) The unit ball in $\mathbb{R}^d$ can be written as

$$\bigcap_{\|w\|=1} \{x : w \cdot x \leq 1\}.$$

3. P_1 and P_2 are polyhedra that are intersections of finitely many halfspaces. Let the halfspaces for P_1 be given by $u_1, \ldots, u_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$:

$$P_1 = \bigcap_{i=1}^{m} \{x : u_i \cdot x \leq b_i\}.$$

Likewise, let P_2 be given by $v_1, \ldots, v_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$:

$$P_2 = \bigcap_{i=1}^{n} \{x : v_i \cdot x \leq c_i\}.$$

We wish to find the point $x_1 \in P_1$ and $x_2 \in P_2$ that are closest to one another. Let us write $z = x_1 - x_2$. Here is the optimization problem:

$$\begin{aligned} &\min \; \|z\|^2 \\ u_i \cdot x_1 &\leq b_i, \quad i = 1, 2, \ldots, m \\ v_i \cdot x_2 &\leq c_i, \quad i = 1, 2, \ldots, n \\ z &= x_1 - x_2 \end{aligned}$$

The constraints are all linear, and the objective function is convex, so this is a convex optimization problem.

4. *Monotone disjunctions.*

(a) There are as many disjunctions as there are subsets of features, so $|\mathcal{H}| = 2^d$.

(b) The true error of h can be bounded thus, with probability at least $1 - \delta$:

$$\text{err}(h) \leq \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta} = \frac{1}{n} \left(d \ln 2 + \ln \frac{1}{\delta} \right).$$

(c) $|\mathcal{H}_k| \leq d^k$, so we get

$$\text{err}(h) \leq \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta} = \frac{1}{n} \left(k \ln d + \ln \frac{1}{\delta} \right).$$

5. By the central limit theorem, $\widehat{p}$ follows roughly a $N(3/4, 1/1600)$ distribution. With 95% probability, $\widehat{p}$ will fall within 2 standard deviations of its mean, that is, in the interval $[0.7, 0.8]$.

6. *VC dimension.*

(a) The class $\mathcal{H}$ of intervals on the real line shatters any set of two distinct points: it can realize all four labelings of these points. But it cannot shatter any set of three points, because it cannot label the middle one 0 while making the other two 1. Therefore $\text{VC}(\mathcal{H}) = 2$.

(b) The class $\mathcal{H}$ of axis-aligned rectangles in the plane shatters the set $\{(0, 1), (0, -1), (1, 0), (-1, 0)\}$: all 16 labelings can be realized. But it cannot shatter any set of five points. To see this, pick any $x_1, \ldots, x_5 \in \mathbb{R}^2$. One of them must lie in the bounding box of the other four points; say x_5 lies in the bounding box of x_1, x_2, x_3, x_4. Then we cannot realize the labeling $y_1 = y_2 = y_3 = y_4 = 1$ and $y_5 = 0$. Thus $\text{VC}(\mathcal{H}) = 4$.

7. *Isotonic regression.*

(a) Here's a monotonic function that goes through four of the points.

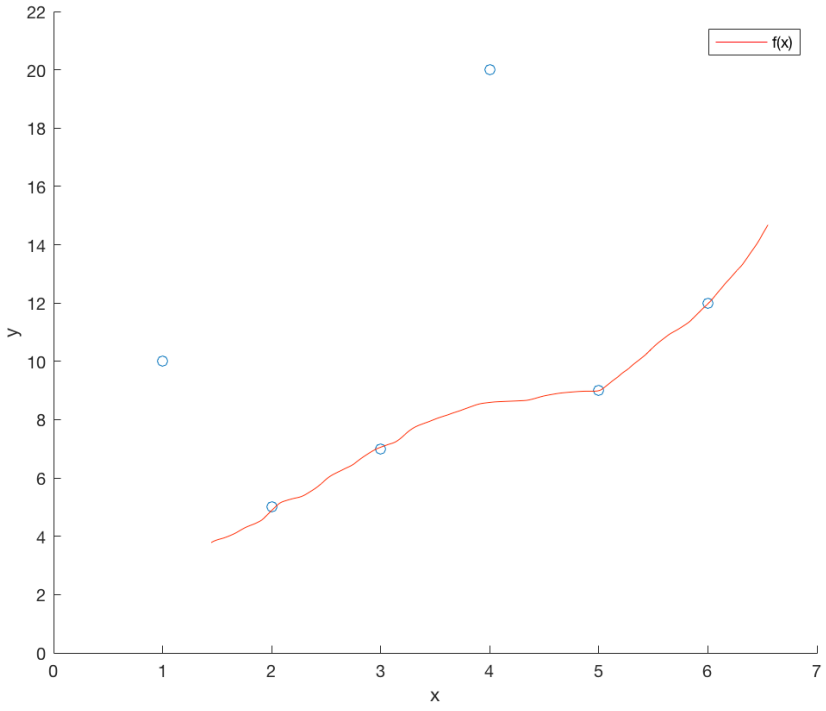


Figure 1: Sketch of $f(x)$

(b) We can write the least-squares isotonic regression problem as follows:

$$\min \; L(f) = \sum_{i=1}^{n} (y_i - f_i)^2$$
$$f_i - f_{i+1} \leq 0 \quad \text{for } i = 1, 2, \ldots, n-1$$

The constraints are linear in f, and the objective function $L(f)$ is convex: its Hessian is $H(f) = 2I$, which is positive semidefinite. Therefore, the problem above is a convex problem.

(c) When the pool-adjacent-violators algorithm is applied to the given set of six points, the final adjusted values are:

$$(1, 22/3), \; (2, 22/3), \; (3, 22/3), \; (4, 41/3), (5, 41/3), (6, 41/3).$$