# Introduction to Learning and Analysis of Big Data HW 1

Tomer Elgavish, Shay Kricheli

November 2019

## 1 Question 2

### 1.1 Section a

The following plot details the mean of the prediction error across 10 iterations as a function of the sample size:

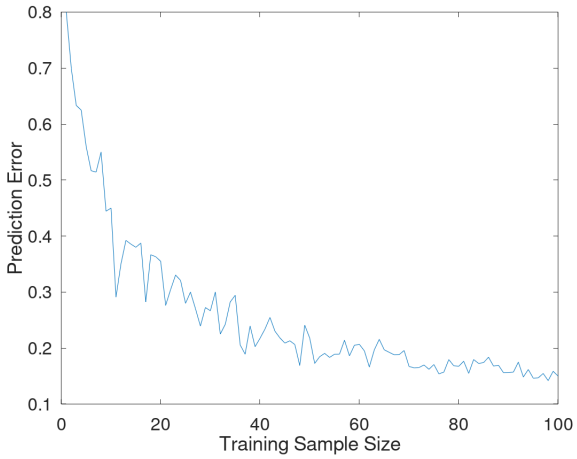


Figure 1: Q2.a - Prediction Error vs. Training Sample Size

### 1.2 Section b

When running the sample size several times - we get different results - as in the error is different. The reason for this is that in different runs we get different samples with different values for the features and therefore the distances between the test and train samples varies - corresponding to different predictions and as a result - different errors.

### 1.3 Section c

We observe that when increasing the sample size - the mean error decreases. This observation is due to the fact that the bigger the training sample we train with - as in containing more samples to compare distances with - the more accurate our prediction will be.

## 1.4 Section d

The following plot details the mean of the prediction error across 10 iterations as a function of the $k$ parameter - which is the number of neighbours of the k-nn algorithm:

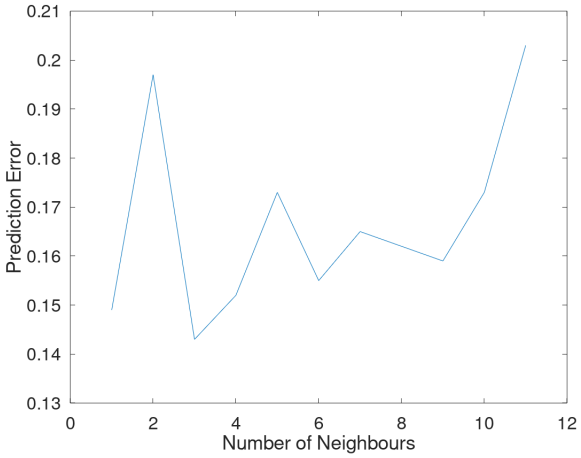


Figure 2: Q2.d - Prediction Error vs. Number of Neighbours

## 1.5 Section e

The plot exhibits an increasing trend in the prediction error when increasing $k$. That behavior is because the sample size is relatively small and when increasing $k$ - we take into consideration neighbours who are further away, who's labels don't necessarily represent the correct prediction label to be considered. If the sample size was increasing along with the $k$ parameter - then we would expect a decreasing trend in the prediction error when increasing $k$.

## 1.6 Section f

The following table details the *confusion matrix* - as defined in the exercise.

| | **0** | **3** | **5** | **8** |
|---|---|---|---|---|
| **0** | 21 | 0 | 1 | 1 |
| **3** | 0 | 25 | 1 | 0 |
| **5** | 1 | 5 | 17 | 2 |
| **5** | 1 | 1 | 3 | 21 |

Table 1: Q2.f Confusion Matrix

The table demonstrates the predictions made in a run for $k = 4$ - the value which had the lowest prediction error in the last section. Each item $(i, j)$ in the table corresponds to the number of predictions which are labeled as $i$ and were predicted to be $j$. The diagonal values correspond to the amount of correct predictions made. The error for this rum was 16% - as one can see is the sum of all off-diagonal elements in the table. This is the case due to the fact that this sum correlates to the number of test examples that were incorrectly predicted.

# 2 Question 3

Let $\mathcal{Y} = \{0, 1\}$ be the set of labels for the following question.

## 2.1 Section a

Let $(X_1, Y_1), (X_2, Y_2)$ be two random labeled examples drawn independently from some distribution $\mathcal{D}$ over $X \times Y$. Suppose that for some examples $x_1, x_2 \in X$, we have $\eta(x_1) = \alpha, \eta(x_2) = \beta$, where: $\eta(x) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = x]$ and $\alpha, \beta \in [0, 1]$. Then:

$$\mathbb{P}[Y_1 \neq Y_2 \mid X_1 = x_1 \wedge X_2 = x_2] =$$
$$\mathbb{P}[(Y_1 = 1 \wedge Y_2 = 0) \vee (Y_1 = 0 \wedge Y_2 = 1) \mid X_1 = x_1 \wedge X_2 = x_2] =$$
$$\mathbb{P}[((Y_1 = 1 \mid X_1 = x_1) \wedge (Y_2 = 0 \mid X_2 = x_2)) \vee ((Y_1 = 0 \mid X_1 = x_1) \wedge (Y_2 = 1 \mid X_2 = x_2))] =$$
$$\mathbb{P}[Y_1 = 1 \mid X_1 = x_1] \cdot \mathbb{P}[Y_2 = 0 \mid X_2 = x_2] + \mathbb{P}[Y_1 = 0 \mid X_1 = x_1] \cdot \mathbb{P}[Y_2 = 1 \mid X_2 = x_2] =$$
$$\eta(x_1) \cdot (1 - \eta(x_2)) + (1 - \eta(x_1)) \cdot \eta(x_2) =$$
$$\alpha(1 - \beta) + \beta(1 - \alpha)$$

Let us explain each transition between lines:

- $1 \to 2$: If $Y_1 \neq Y_2$, then since $Y_1, Y_2 \in \mathcal{Y} = \{0, 1\}$ then that means that either $Y_1 = 1$ and $Y_2 = 0$ or $Y_1 = 0$ and $Y_2 = 1$.
- $2 \to 3$: Since $(X_1, Y_1)$ is drawn independently from $(X_2, Y_2)$, their corresponding conditional probabilities are also independant.
- $3 \to 4$: The events $Y_1 = 1$ and $Y_2 = 0$ are disjoint.
- $4 \to 5$: Using the definition of $\eta(x)$.
- $5 \to 6$: Using the definition of $\alpha$ and $\beta$.

## 2.2 Section b

Let $\mathcal{D}$ be some distribution over $X \times Y$ such that for all $x \in X$, $\eta(x) = \alpha$ for some $\alpha \in [0,1]$. Let us calculate the Bayes-optimal error of $\mathcal{D}$:

$$
\begin{aligned}
&err(h^*, \mathcal{D}) = \sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot (1 - \mathbb{P}[Y = h^*(x) \mid X = x]) = \\
&\sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot (1 - \mathbb{P}[Y = \mathbb{I}[\eta(x) > 1/2] \mid X = x]) = \\
&\sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot (1 - \mathbb{P}[Y = \mathbb{I}[\alpha > 1/2] \mid X = x]) = \\
&\sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot \begin{cases} 1 - \mathbb{P}[Y = 1 \mid X = x] & ;\alpha > 1/2 \\ 1 - \mathbb{P}[Y = 0 \mid X = x] & ;\alpha \leq 1/2 \end{cases} = \\
&\sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot \begin{cases} 1 - \eta(x) & ;\alpha > 1/2 \\ 1 - (1 - \eta(x)) & ;\alpha \leq 1/2 \end{cases} = \sum_{x \in \mathcal{X}} \mathbb{P}[X = x] \cdot \begin{cases} 1 - \alpha & ;\alpha > 1/2 \\ \alpha & ;\alpha \leq 1/2 \end{cases} = \\
&\begin{cases} 1 - \alpha & ;\alpha > 1/2 \\ \alpha & ;\alpha \leq 1/2 \end{cases} \cdot \sum_{x \in \mathcal{X}} \mathbb{P}[X = x] = \begin{cases} 1 - \alpha & ;\alpha > 1/2 \\ \alpha & ;\alpha \leq 1/2 \end{cases} = \min\{\alpha, 1 - \alpha\}
\end{aligned}
$$

Let us explain:

- Line 1: By the definition of the Bayes-optimal error as seen in class.
- $1 \to 2$: Since we are using the Bayes-optimal rule, we have $h^*(x) := \mathbb{I}[\eta(x) > 1/2]$.
- $2 \to 3$: Using the definition of $\alpha$.
- $3 \to 4$: Splitting into cases by the definition of the indicator function $\mathbb{I}$ corresponding to the value of $\alpha$.
- $4 \to 5$: Using the definition of $\eta(x)$ and again $\alpha$.
- $5 \to 6$: The value of $\alpha$ is constant for all $x \in \mathcal{X}$ and so is independant of $x$.
  The sum $\sum_{x \in \mathcal{X}} \mathbb{P}[X = x]$ corresponds to the probability of drawing some example out of all of the examples - which is of course 1. Finally, since the best prediction rule is the Bayes-optimal rule, the expression for the error is achieved when taking the minimum of the two cases where $\alpha > 1/2$ and $\alpha \leq 1/2$.

## 2.3 Section c

Suppose that $\mathcal{X} = [0,1]$ for some distribution $\mathcal{D}$ with the same property of $\eta$ as defined in the last section, and that the marginal of $\mathcal{D}$ on $\mathcal{X}$ in uniform. Let $S \sim \mathcal{D}^m$ be a sample set and $x \in \mathcal{X}$ be some example drawn. Let:

$$\begin{gathered}\mathbb{P}_{S\sim\mathcal{D}^m}[\exists y \in \mathcal{Y} \; s.t. \; (x,y) \in S] = \\ \mathbb{P}_{S=\cup_{i=1}^m \{(x_i,y_i)\}\sim\mathcal{D}^m}[\cup_{i=1}^m x_i = x] \leq \\ \sum_{(x_i,y_i)\in S} \mathbb{P}[x_i = x] = m \cdot 0 = 0 \; \blacksquare\end{gathered}$$

Let us explain:

- $1 \rightarrow 2$: The expression in line 1 refers to the probability that some pair $(x_i, y_i) \in S$ holds $x_i = x$ and so that expression can be converted into a union of $m$ events - one for each pair in $S$.
- $2 \rightarrow 3$: Applying the union bound; for a set of events $\{A_i\}_{i=1}^n$: $\mathbb{P}[\cup_{i=1}^n A_i] \leq \sum_{i=1}^n \mathbb{P}[A_i]$.
- Line 3: The probability that a certain number $x_i$ from a pair $(x_i, y_i) \in S$ equals to the number $x$ drawn from the continuous set $[0,1]$ is zero.

Let the probability that some example is repeated more than once in $S$ be:

$$\begin{gathered}\mathbb{P}_{S\sim\mathcal{D}^m}[\exists y_1, y_2 \in \mathcal{Y}, \exists x \in \mathcal{X} \; s.t. \; (x,y_1),(x,y_2) \in S] = \\ \mathbb{P}_{S=\cup_{1\leq i<j\leq m}\{(x_i,y_i),(x_j,y_j)\}\sim\mathcal{D}^m}[\cup_{1\leq i<j\leq m} x_i = x_j] \leq \\ \sum_{(x_i,y_i)\in S} \sum_{\substack{(x_j,y_j)\in S \\ j\neq i}} \mathbb{P}[x_i = x_j] = \binom{m}{2} \cdot 0 = 0 \; \blacksquare\end{gathered}$$

In this equation, similar to was done above - we went over all of the possible pairs $(x_i, y_i), (x_j, y_j) \in S$ and used the union bound again to end up with the expression for the probability that $x_i = x_j$ - which is zero as was discussed.

## 2.4 Section d

In this section we are to calculate the expected error of the nearest-neighbor algorithm for the distribution $\mathcal{D}$. Let $S = \{(x_1, y_1), ..., (x_m, y_m)\} \sim \mathcal{D}^m$ be the training sample and let $x \in \mathcal{X}$ be an additional example such that $x, x_1, ..., x_m$ are different from each other. For $(X, Y) \in \mathcal{D}$, let:

$$\begin{gathered}f(\alpha) = \mathbb{P}_{S\sim\mathcal{D}^m}[Y \neq h(X) \mid X = x \wedge (x_1,y_1), ..., (x_m,y_m) \in S] = \\ \mathbb{P}_{S\sim\mathcal{D}^m}[Y \neq y_{nn(x)} \mid X = x \wedge (x_1,y_1), ..., (nn(x),y_{nn(x)}), ..., (x_m,y_m) \in S] = \\ \eta(nn(x))(1-\eta(x)) + \eta(x)(1-\eta(nn(x))) = \alpha(1-\alpha) + \alpha(1-\alpha) = 2\alpha(1-\alpha)\end{gathered}$$

Let us explain:

- $1 \rightarrow 2$: As explained in the question description, once we know the examples $x_1, ..., x_m \in \mathcal{X}$ and the example $x \in \mathcal{X}$ for which the prediction is required, we can tell which of the examples in the sample is the nearest neighbor of $x$ - which we'll denote as $nn(x)$. Moreover, we are using the nearest-neighbour algorithm so our prediction rule is: $h(x) = y_{nn(x)}$.
- $2 \rightarrow 3$: Applying the result achieved in section a to $x$ and $nn(x)$.

## 2.5 Section e

In this section we are asked to plot the expression for $\mathbb{E}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D})]$ and the Bayes-optimal error for values of $\alpha \in [0, 1]$. As mentioned in the previous section description: $\mathbb{E}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D})] = f(\alpha)$. Let us then plot $f(\alpha)$ and the Bayes-optimal error:

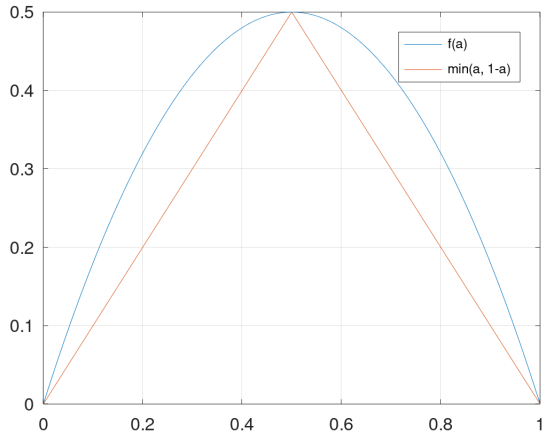


Figure 3: Q3.e - Bayes-optimal error and $\mathbb{E}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D})]$

## 2.6 Section f

Let $f(\alpha) = 2\alpha(1-\alpha)$, $e(\alpha) = \min\{\alpha, 1-\alpha\}$, $h(\alpha) = 2e(\alpha) = \min\{\alpha, 1-\alpha\}$. We'll show that $e(\alpha) \leq f(\alpha) \leq h(\alpha)$ for all $\alpha \in [0, 1]$. We'll discuss four cases:

- $0 < \alpha < 1/2$:
  In that case since $\alpha < 1-\alpha$, we'll get $e(\alpha) = \alpha$. So $h(\alpha) = 2\alpha$. $\alpha > 0$ so $1-\alpha < 1$ so $f(\alpha) = 2\alpha(1-\alpha) < 2\alpha = h(\alpha)$. So $f(\alpha) < h(\alpha)$. Moreover, $\alpha < 1/2$ so $1-\alpha > 1/2$ so $f(\alpha) = 2\alpha(1-\alpha) > 2\alpha \cdot 1/2 = \alpha = e(\alpha)$ and then we get $f(\alpha) > e(\alpha)$.

- $1/2 < \alpha < 1$:
  In that case since $\alpha > 1-\alpha$, we'll get $e(\alpha) = 1-\alpha$. So $h(\alpha) = 2(1-\alpha)$. $\alpha < 1$ so $f(\alpha) = 2\alpha(1-\alpha) < 2(1-\alpha) = h(\alpha)$. So $f(\alpha) < h(\alpha)$. Moreover, $\alpha > 1/2$ so $f(\alpha) = 2\alpha(1-\alpha) > 2 \cdot 1/2 \cdot (1-\alpha) = (1-\alpha) = e(\alpha)$ and then we get $f(\alpha) > e(\alpha)$.

- $\alpha = 1/2$:
  In that case we'll get $f(\alpha) = e(\alpha) = 1/2$ - in that case $f(\alpha)$ gets the same value as the Bayes-optimal error.

- $\alpha = 0$ or $\alpha = 1$:
  In that case we'll get $f(\alpha) = e(\alpha) = h(\alpha) = 0$ - in that case the error gets the same value as twice the Bayes-optimal error.

# 3 Question 4

## 3.1 Section a

Suppose that $a \in [\beta - \epsilon, \beta + \epsilon]$. We will split it into two possible foreign cases, and show that for each case, $err(f_a, \mathcal{D}) \leq \epsilon$:

1. $\boldsymbol{a \in [\beta - \epsilon, \beta]}$. Let's assume that $(X, Y) \sim \mathcal{D}$. The case in which $f_a$ will give a different label to $X$, can only happen if $x \in [a, \beta]$ - in this case $Y = 0$ because $x \leq \beta$, but $f_a(X) = 1$ because $x \geq a$. Therefore, we can write:
$\mathbb{P}_{(X,Y)\sim\mathcal{D}}[f_a(X) \neq Y] = \mathbb{P}[x \in [a, \beta]]$.
Now, this probability will be the biggest when $a = \beta - \epsilon$ because $x$ will have the most options to be chosen from. So:
$\mathbb{P}[x \in [a, \beta]] \leq \mathbb{P}[x \in [\beta - \epsilon, \beta]] = \epsilon$. Therefore, $\mathbb{P}_{(X,Y)\sim\mathcal{D}}[f_a(X) \neq Y] \leq \epsilon$.

2. $\boldsymbol{a \in [\beta, \beta + \epsilon]}$. For the same reasons as the first case, we can write:
$\mathbb{P}_{(X,Y)\sim\mathcal{D}}[f_a(X) \neq Y] = \mathbb{P}[x \in [\beta, a]] \leq \mathbb{P}[x \in [\beta, \beta + \epsilon]] = \epsilon$.

## 3.2 Section b

Suppose there are $(x_1, y_1), (x_2, y_2) \in S$ such that $x_1 \in [\beta - \epsilon, \beta], x_2 \in [\beta, \beta + \epsilon]$.
Suppose negatively that $\hat{h}_S = f_a$ for $a > \beta + \epsilon$ or $a < \beta - \epsilon$.
In both cases, $\mathbb{P}_{(X,Y)\sim S}[\hat{h}_S(X) \neq Y] > 0$ because if $a > \beta + \epsilon$ then $\hat{h}_S(x_2) \neq y_2$, and if $a < \beta - \epsilon$ then $\hat{h}_S(x_1) \neq y_1$.
Therefore, $err(\hat{h}_S, S) = \mathbb{P}_{(X,Y)\sim S}[\hat{h}_S(X) \neq Y] > 0$.
Since for each $x \in X$, $\mathbb{P}[Y = 1|X = x] = \mathbb{I}[x \geq \beta]$, we will infer that for the Bayes-optimal predictor, $err(f_\beta, S) = 0$.
Since $\hat{h}_S$ is the output classifier returned by an ERM algorithm, we get a contradiction!
Therefore $\hat{h}_S = f_a$ for $a \in [\beta - \epsilon, \beta + \epsilon]$.
From the previous section 3.*a*, since $a \in [\beta - \epsilon, \beta + \epsilon]$, we conclude that $err(f_a, D) \leq \epsilon$.

## 3.3 Section c

$\mathbb{P}_{S\sim\mathcal{D}^m}[\nexists(x, y) \in S s.t. x \in [\beta, \beta + \epsilon]] = \mathbb{P}_{S\sim\mathcal{D}^m}[\forall(x, y) \in S : x \notin [\beta, \beta + \epsilon]] = \mathbb{P}_{S=\cup_{i=1}^m (x_i,y_i)\sim\mathcal{D}^m}[x_1 \notin [\beta, \beta+\epsilon] \wedge x_2 \notin [\beta, \beta+\epsilon] \wedge ... \wedge x_m \notin [\beta, \beta+\epsilon]] = \prod_{i=1}^m \mathbb{P}[x_i \notin [\beta, \beta + \epsilon]] = \prod_{i=1}^m 1 - \mathbb{P}[x_i \in [\beta, \beta + \epsilon]] = \prod_{i=1}^m (1 - \epsilon) = (1 - \epsilon)^m$

The events $x_i \notin [\beta, \beta + \epsilon]$ are independent for each $1 \leq i \leq m$, and therefore the transition between the third expression and the fourth expression is legal.

The same holds for the existence of $x \in [\beta - \epsilon, \beta]$ because $\mathbb{P}[x_i \in [\beta - \epsilon, \beta]]$ is still $(1 - \epsilon)$ as before.

From subsection 3.*a*, we can infer that if $err(f_a, D) > \epsilon$ then $a \notin [\beta - \epsilon, \beta + \epsilon]$. Therefore:
$\mathbb{P}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D}) \leq \epsilon] = 1 - \mathbb{P}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D}) > \epsilon] = 1 - \mathbb{P}[a \notin [\beta - \epsilon, \beta + \epsilon]] = 1 - \mathbb{P}[a < (\beta - \epsilon) \vee a > (\beta + \epsilon)] \geq 1 - (\mathbb{P}[a < \beta - \epsilon] + \mathbb{P}[a > \beta + \epsilon]) \geq 1 - (\mathbb{P}[\nexists(x, y) \in S s.t. x \in [\beta - \epsilon, \beta]] + \mathbb{P}[\nexists(x, y) \in S s.t. x \in [\beta, \beta + \epsilon]]) = 1 - 2(1 - \epsilon)^m$

### 3.4 Section d

In order to guarantee that there is a probability of at least 95 percent that the output of the ERM algorithm has an error of at most 3 percent on the distribution, we need a sample size of $m = 122$. That is because, if we modelling the question into the above inequality, we need:
$\mathbb{P}_{S\sim\mathcal{D}^m}[err(\hat{h}_S, \mathcal{D}) \leq 0.03] \geq 0.95$.
Therefore, we are looking to calculate $1 - 2 \cdot (1 - 0.03)^m \leq 0.95$:
$1 - 2 \cdot (1 - 0.03)^m \leq 0.95$
$2 \cdot 0.97^m \geq 0.05$
$0.97^m \geq 0.025$
$m \geq 121.1$

## 4 Question 5

### 4.1 Section a

The set of all possible examples - denoted $\mathcal{X}$ in this problem is a set containing the two features that define the patients - their height in cm and their age in years. Thus $\mathcal{X}$ is a set of 2 dimensional vectors over $\mathbb{R}$; $\mathcal{X} = \cup_i\{(h_i, a_i) \mid 0 \leq h_i \leq 250 \wedge 0 \leq a_i \leq 120\} \subseteq \mathbb{R}^2$.
The the set of all possible labels - denoted $\mathcal{Y}$ in this problem is the two possible treatments defined in the problem for the patients - take medicine and rest at home. Let us define: $\mathcal{Y} = \{0, 1\}$ where 0 stands for take medicine and 1 stands for rest at home.

### 4.2 Section b

The Bayes-optimal predictor $h^*$ for distribution $\mathcal{D}$ is, by definition:
$h^*(x) = \mathbb{I}[\eta(x) > 1/2]$ , where $\eta(x) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = x]$ and $\mathcal{Y} = \{0, 1\}$. Next we are to write the value of the predictor $h^*(x)$ for each possible $x$ with a non-zero probability in $\mathcal{D}$. Let us look of an example of calculation:

$$h^*((160, 20)) = \mathbb{I}[\eta((160, 20)) > 1/2] = \mathbb{I}[\mathbb{P}_{(X,Y)\sim\mathcal{D}}[Y = 1 \mid X = (160, 20)] > 1/2] = \mathbb{I}[1 > 1/2] = 1$$

Similarly, we have:

$$h^*((160, 40)) = 1, h^*((180, 25)) = 1, h^*((180, 35)) = 0$$

The Bayes-optimal error of $\mathcal{D}$ is:

$$err(h^*, \mathcal{D}) = \sum_{x\in X} \mathbb{P}[X = x] \cdot [1 - \mathbb{P}[Y = h^*(x)|X = x]] =$$

$$0.1 \cdot (1 - 1) + 0.5 \cdot (1 - 0.9) + 0.25 \cdot (1 - \frac{2}{3}) + 0.15 \cdot (1 - 1) = \frac{2}{15}$$

## 4.3 Section c

In this section we are asked to find a linear predictor (with a bias) that gives the Bayes-optimal error on $\mathcal{D}$. The predictor $w = (-2/3, -1)$ , $b = 200$ is a suiting linear predictor for this problem. As seen in class, the prediction rule will be: $h_{w,b}(x) = sign(\langle w, x\rangle + b) = sign(\langle(-2/3, -1), x\rangle + 200)$, where a $+1$ sign will correspond to the label 1 and the $-1$ sign will correspond to the label 0. The following plot demonstrates the predictor with the scatter points corresponding to the patients' features:

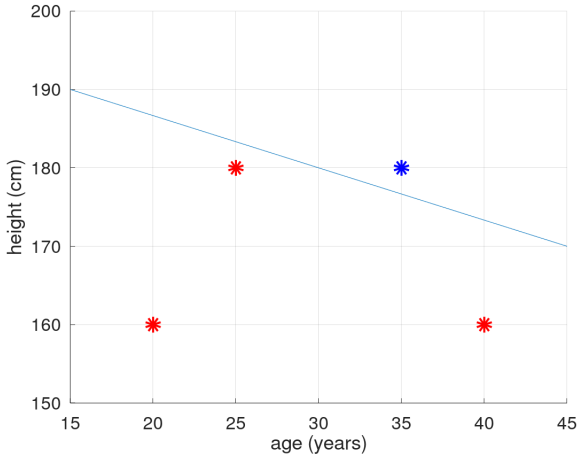


Figure 4: Q5.c - Linear Predictor

To show that the predictor results in the Bayes-optimal error, we'll show that it's equal to $h^*(x)$ - the Bayes-optimal rule:

$$h_{w,b}((160, 20)) = sign(\langle(-2/3, -1), (160, 20)\rangle + 200) = sign(93.33) = 1 = h^*((160, 20))$$
$$h_{w,b}((160, 40)) = sign(\langle(-2/3, -1), (160, 40)\rangle + 200) = 1 = h^*((160, 40))$$
$$h_{w,b}((180, 25)) = sign(\langle(-2/3, -1), (180, 25)\rangle + 200) = 1 = h^*((180, 25))$$
$$h_{w,b}((180, 35)) = sign(\langle(-2/3, -1), (180, 35)\rangle + 200) = 0 = h^*((180, 35))$$

Thus we have that our linear predictor predicts the same results as the Bayes-optimal prediction rule, and thus:

$$err(h_{w,b}, \mathcal{D}) = err(h^*, \mathcal{D}) = \frac{2}{15}$$

## 4.4 Section d

| height | best treatment | probability |
|---|---|---|
| 160 | rest | 10% |
| 160 | rest | 45% |
| 160 | medicine | 5% |
| 180 | rest | 15% |
| 180 | medicine | 10% |
| 180 | medicine | 15% |

The Bayes-optimal error for this distribution is:
$err(h^*, \mathcal{D}^*) = \sum_{x \in X} \mathbb{P}[X = x] \cdot [1 - \mathbb{P}[Y = h^*(x)|X = x]] = 0.6 \cdot (1 - \frac{11}{12}) + 0.4 \cdot (1 - \frac{5}{8}) = 0.2.$

### 4.5 Section e

In this section we are given a different distribution - $\mathcal{D}'$. The Bayes-optimal predictor for $\mathcal{D}'$ is:

$$h^*((160, 20)) = 0, h^*((170, 20)) = 1, h^*((180, 20)) = 0$$

The Bayes-optimal error in this case is also 0, but it cannot be achieved by a linear predictor because, as explained in class, this distribution is inseparable, as can be seen in the following plot:

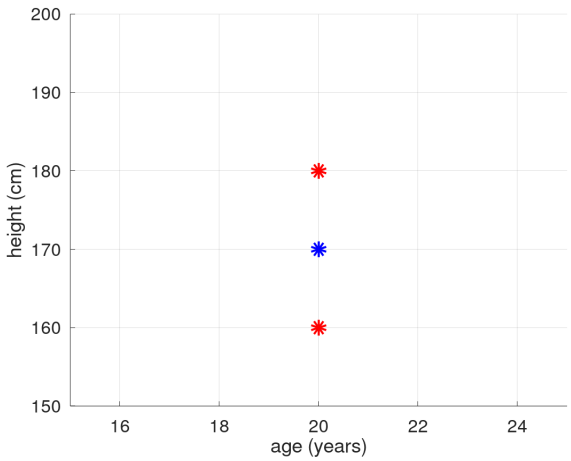


Figure 5: Q5.c Distribution Scatter