CS/ECE/ME532 Period 2 Activity

1) Let $\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & -2 & -2 \\ 3 & -3 & 3 & -3 \end{bmatrix}$ and $\boldsymbol{w} = \begin{bmatrix} 1 \\ b \\ 1 \\ c \end{bmatrix}$.

   a) Write out and evaluate the vector $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$.

   b) Find b and c so that $\boldsymbol{y} = \begin{bmatrix} 4 \\ 0 \\ 0 \end{bmatrix}$.

   c) Find b and c so that $\boldsymbol{y} = \begin{bmatrix} 0 \\ 0 \\ 12 \end{bmatrix}$.

SOLUTION:

a) $\boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & -2 & -2 \\ 3 & -3 & 3 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ b \\ 1 \\ c \end{bmatrix} = \begin{bmatrix} 2+b+c \\ 2(b-c) \\ 6-3(b+c) \end{bmatrix}$

b) Note that $b - c = 0$ so $b = c$. Then $2 + b + c = 4$ implies $b = c = 1$. This also satisfies $6 - 3(b + c) = 0$.

c) Again we require $b = c$. Using $2 + b + c = 0$ we have $b = c = -1$. This also satisfies $6 - 3(b + c) = 12$.

2) Recall from the last activity that food involves fats, proteins and carbohydrates. There are 9 calories for every gram of fat, 4 calories for every gram of protein, and 4 calories for every gram of carbohydrates. If we define a vector $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ where x_1 is the number of grams of fat, x_2 is the number of grams of protein, and x_3 is the number of grams of carbohydrate, then the number of calories is $y = \boldsymbol{x}^T\boldsymbol{w}$ where $\boldsymbol{w} = \begin{bmatrix} 9 \\ 4 \\ 4 \end{bmatrix}$.

Your nutrition expert has a way of defining a food as "low carb" based on the ratio of carbohydrate calories to total calories. Let

$$z = \frac{\text{carbohydrate calories}}{\text{total calories}}$$

A food is classified as low carb if $z < 1/4$.

a) Express the rule for classifying foods given by your nutritionist as the sign of an inner product between $\boldsymbol{x}$ and a vector of weights $\tilde{\boldsymbol{w}}$. In other words, specify $\tilde{w}_1, \tilde{w}_2$ and $\tilde{w}_3$ so that when $\text{sign}(\boldsymbol{x}^T\tilde{\boldsymbol{w}}) = 1$ then the food is low carb, and when $\text{sign}(\boldsymbol{x}^T\tilde{\boldsymbol{w}}) = -1$ then the food is not low carb.

b) Nutritionists like to look at the ratios of the types of calories. Consider the features $r_f = x_1/x_3$, the ratio of the number of grams of fat to carbohydrate, and $r_p = x_2/x_3$, the ratio of the number of grams of protein to carbohydrate. Express low carb criterion as a function of the features r_f and r_p.

c) Define the decision boundary as the line where $z = 1/4$, since a food with $z < 1/4$ is classified as low carb while a food with $z \geq 1/4$ is not low carb. Graph the decision boundary assuming feature r_p is on the vertical axis and r_f is on the horizontal axis. Shade the portion of the r_f-r_p plane that corresponds to low-carb foods. Note that r_p and r_f cannot be negative. *Hint:* Recall that the equation $y = mx + b$ describes a line with slope m and y-intercept b.

d) Consider the four cereals:

Cereal 1: 1 gram fat, 8 grams protein, 44 grams carbohydrate
Cereal 2: 0.5 grams fat, 2 grams protein, 25 grams carbohydrate
Cereal 3: 1.3 grams fat, 2.7 grams protein, 29.3 grams carbohydrate
Cereal 4: 9 grams fat, 4 grams protein, 16 grams carbohydrate

Plot the features r_f, r_p for each cereal in the r_f-r_p plane and label each pair of features with the corresponding cereal number. Are any of these classified as low carb?

e) Almond butter has 9 grams fat, 3.4 grams protein, and 3 grams carbohydrate per serving. Plot the features r_f, r_p for almond butter in the r_f-r_p plane. Is almond butter classified as a low-carb food?

f) A serving of marinated grilled salmon has 19 grams fat, 23 grams protein, and 1 gram carbohydrate per serving. Plot the features r_f, r_p in the r_f-r_p plane. Is this salmon classified as a low-carb food?

SOLUTION:

a) Low carb if

$$\frac{9}{4}x_1 + x_2 - 3x_3 > 0$$

hence $\tilde{w}_1 = 9/4$, $\tilde{w}_2 = 1$, $\tilde{w}_3 = -3$, or some scaling.

b) Low carb if

$$\frac{9}{4}r_f + r_p - 3 > 0$$

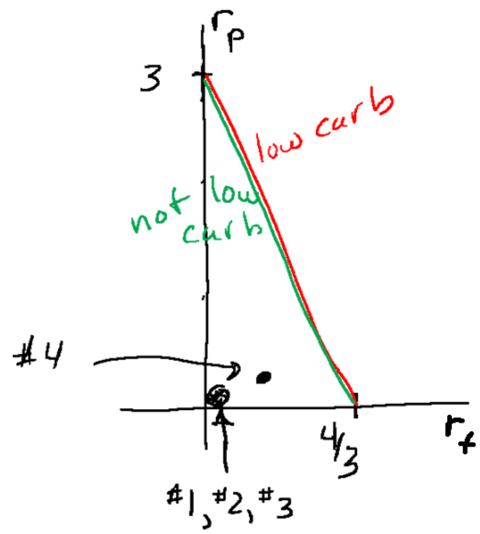
c) $4 = \frac{9}{4}r_f + r_p + 1$ implies that $r_p = -\frac{9}{4}r_f + 3$. This is the equation for a line with slope $-\frac{9}{4}$ and vertical intercept $r_p = 3, r_f = 0$

d) Features:

Cereal 1: $r_f = 0.023, r_p = 0.18$
Cereal 2: $r_f = 0.02, r_p = 0.08$
Cereal 3: $r_f = 0.044, r_p = 0.092$
Cereal 4: $r_f = 0.563, r_p = 0.25$

None of these are classified as low carb.

e) Almond butter: $r_f = 3, r_p = 1.13$ is low carb

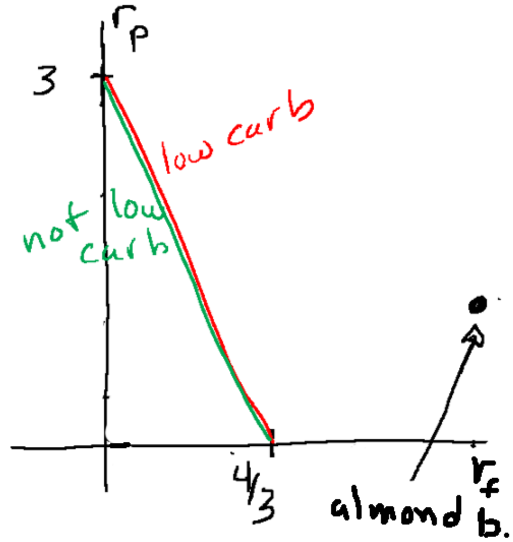


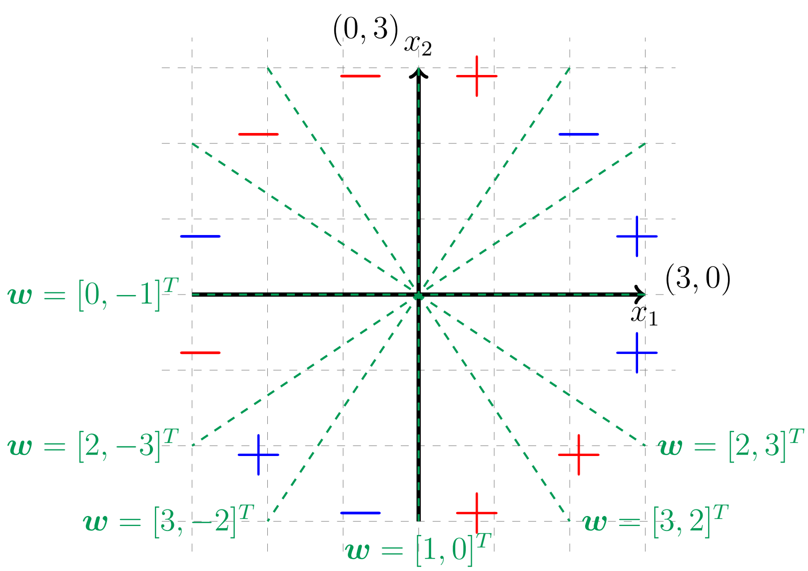
f) Salmon: $r_f = 19, r_p = 23$ is low carb and off the scale on the sketch.

Fairness in machine learning. Various forms of discrimination could be built into the machine learning algorithms. You can find more details from these articles: 1) "Artificial Intelligence's White Guy Problem" by NY Times and 2) "Machine Bias" by Propublica. This led the researchers to introduce the notion of "fair machine learning", which we will touch upon in the following problem. Ensuring fairness in machine learning is a current research topic, so please refer to "Fairness and machine learning: Limitations and Opportunities" by Barocas, Hardt, and Narayanan for an in-depth survey.

3) Decision boundaries that lead to the best overall performance may result in poor performance for subgroups of the data that have slightly different characteristics.

Let $\boldsymbol{x}^T = [x_1, x_2]$. Consider the following dataset.





a) Consider the following linear classifier:

$$\boldsymbol{x}^T\boldsymbol{w} \geq 0 : \text{predict } +,$$
$$\boldsymbol{x}^T\boldsymbol{w} < 0 : \text{predict } -,$$

where $\boldsymbol{w}$ is one of the 6 vectors and corresponding decision boundaries shown above. For example, $\boldsymbol{w} = [1,0]^T$ results in the classifier

$$x_1 \geq 0 : \text{predict } +,$$
$$x_1 < 0 : \text{predict } -,$$

so values to the right of the x_2-axis are classified as $+$ and those to the left as $-$. Which decision boundaries in the picture (i.e, which $\boldsymbol{w}$'s) minimize the total number of misclassifications?

SOLUTION: $\boldsymbol{w} = [1, 0]^T$ and $\boldsymbol{w} = [2, -3]^T$.

b) A classifier is called *fair* if it performs equally well on specified subgroups of the dataset, i.e., the percentage of misclassifications for each subgroup is the same. The above dataset consists of two subgroups: six blue data points and six red data points. Of the six candidate classifiers, find the ones that are fair.

SOLUTION: By inspection, we can see that the following $\boldsymbol{w}$'s are fair:

$$\boldsymbol{w} = [0, -1]^T, \boldsymbol{w} = [2, -3]^T, \boldsymbol{w} = [0, -1]^T, \boldsymbol{w} = [3, 2]^T.$$

c) Among all the fair classifiers, find the one that minimizes the total number of misclassification. Are all the classifiers that you found in part a) fair? **SOLUTION:** $\boldsymbol{w} = [2, -3]^T$