CSCI 467 Problem Set 1
January 6, 2020
Teresa Liu

1) Problem 1
    a. Unsupervised, since we want to produce 256 groups of colors but we don't know what specific colors we are looking for from the million distinct colors, so we wish to perform dimensionality reduction.
    b. Semi-supervised, because you have a large dataset of labeled data and you've used a crawler to download another large number of webpages. Since you wish to use both datasets, one labeled, one unlabeled, this falls between supervised and unsupervised learning.
    c. Supervised, assuming that the dataset of photo IDs have both age and portrait. Since you already have labeled data of portraits and ages, we can estimate the age of a person.
    d. Reinforcement learning, since the agent (in this case, the vacuum) needs to navigate an environment with a set of obstacles.
    e. Unsupervised, since we wish to perform dimensionality reduction to go from 200 features to 20 summarized new features.
    f. Supervised, since we already have the labeled data of insurance claims with time processed that we can build a model on top of.
2) Problem 2
    a. Classification, because we are looking for discrete outputs i.e. seabass or salmon.
    b. Regression, since the price of a house is quantitative and not discrete.
    c. Classification, since the outputs are discrete, either has diabetes or does not have diabetes.
    d. Regression, since the MPG of a car is quantitative and not discrete.
    e. Classification, since there's 16 categories of Myers Briggs personalities that would be our discrete outputs.
3) Problem 3
    a. No this statement does not defy the no free lunch theorem. Even though these 5 algorithms perform similarly, when applied to all other problems, they still may follow the logic that the no free lunch theorem states: that averaged over all problems, those algorithms perform no better than other optimization/search algorithms. It is possible and in line with the theorem that these algorithms might perform very different on other data sets to average out their performance over all problems.
    b. No this statement does not defy the no free lunch theorem, because even though the special type of classifier may work well with text data, we are not told that it works well on every problem. Therefore, we can assume according to the free lunch theorem, that the classifier's performance averaged among other problems beyond text data problems does not perform any better than other algorithms.

4) Problem 4
    a. Wf = 0.25, Wv = 0.2, Wd = 0.3, Ws = 0.25

Person	Sim(I,Q) with query (High, No, Yes, Yes)
1, instance (No, No, No, No)	= 0.25(0) + 0.2(1) + 0.3(0) + 0.25(0) = 0.2
2, instance (Low, No, No, No)	= 0.25(0.3) + 0.2(1) + 0.3(0) + 0.25(0) = 0.275
3, instance (High, No, No, Yes)	= 0.25(1) + 0.2(1) + 0.3(0) + 0.25(1) = 0.7
4, instance (High, Yes, Yes, No)	= 0.25(1) + 0.2(0) + 0.3(1) + 0.25(0) = 0.55
5, instance (Low, No, Yes, No)	= 0.25(0.3) + 0.2(1) + 0.3(1) + 0.25(0) = 0.575
6, instance (No, Yes, Yes, No)	= 0.25(0) + 0.2(0) + 0.3(1) + 0.25(0) = 0.3

7, instance (Low, Yes, Yes, No)	= 0.25(0.3) + 0.2(0) + 0.3(1) + 0.25(0) = 0.375

b. With test instance (*, Yes, Yes, No), you can take the average of the Fever attribute for every possible query value of Fever and multiply it by the weight.

Person	Sim(I,Q) with query (*, Yes, Yes, No)
1, instance (No, No, No, No)	= 0.25(1.5/3) + 0.2(0) + 0.3(0) + 0.25(1) = 0.375
2, instance (Low, No, No, No)	= 0.25(2/3) + 0.2(0) + 0.3(0) + 0.25(1) = 0.41667
3, instance (High, No, No, Yes)	= 0.25(2/3) + 0.2(0) + 0.3(0) + 0.25(0) = 0.1667
4, instance (High, Yes, Yes, No)	= 0.25(2/3) + 0.2(1) + 0.3(1) + 0.25(1) = 0.91667
5, instance (Low, No, Yes, No)	= 0.25(2/3) + 0.2(0) + 0.3(1) + 0.25(1) = 0.71667
6, instance (No, Yes, Yes, No)	= 0.25(1.5/3) + 0.2(1) + 0.3(1) + 0.25(1) = 0.875
7, instance (Low, Yes, Yes, No)	= 0.25(2/3) + 0.2(1) + 0.3(1) + 0.25(1) = .91667

c. We must find the largest similarity since distance and similarity are inversely proportional. With k=3,
    i. In 4a) with Query (High, No, Yes, Yes) the closest neighbors are
        1. Person 3, Sim(I,Q) = 0.7
        2. Person 4, Sim(I,Q) = 0.55
        3. Person 5, Sim(I,Q) = 0.575
        4. Diagnosis is Food Poisoning
    ii. In 4b) with Query(*, Yes, Yes, No) the closest neighbors are
        1. Person 4, Sim(I,Q) = 0.91667
        2. Person 7, Sim(I,Q) = 0.91667
        3. Person 6, Sim(I,Q) = 0. 875
        4. Diagnosis is Stomach Flu

5)

a) $p_{W|H}(w|h) = \frac{1}{\sqrt{2\pi} \times 10} \frac{-(w-0.5\times h^{1.001})^2}{200}$

- PDF of Gaussian Distribution = $\frac{1}{\sigma\sqrt{2\pi}} \frac{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}{}$ $w \to x$

therefore $\mu = h^{1.001}$, $\sigma = 10$

- In lecture, we state that the regression function $f(x) = f(x_1, x_2, x_3) = E[Y | X_1 = x_1, X_2 = x_2, X_3 = x_3]$ so in this case, $w = f(h) = E[W|H=h]$
- In Gaussian distribution, $\mu = h^{1.001} = E[W|H=h]$

So $\boxed{w = f(h) = h^{1.001}}$

b) $h = 155$ cm $\quad w = h^{1.001} = (155)^{1.001} = 155.78$ kg

$h = 165$ cm $\quad w = h^{1.001} = (165)^{1.001} = 165.849$ kg

$h = 190$ cm $\quad w = h^{1.001} = (190)^{1.001} = 190.9996$ kg

c) No, because the mean of the Gaussian distribution would still be $h^{1.001}$, and since $E[W|H=h] = \mu = f(h)$ we would have the same answers. The variance does not affect the mean.

d) $h = 150$ cm : nearest : person 5 (150 cm), 2 (168 cm), 1 (171 cm)

$\hat{y}_{KNN} = \frac{65+78+80}{3} = \boxed{74.33 \text{ kg}}$

$h = 155$ cm : nearest : person 5 (150 cm), 2 (168 cm) 1 (171 cm)

$\hat{y}_{KNN} = \frac{65+78+80}{3} = \boxed{74.33 \text{ kg}}$

$h = 165$ cm : nearest : person 2 (168 cm), 1 (171 cm), 6 (178 cm)

$\hat{y}_{KNN} = \frac{78+80+83}{3} = \boxed{80.33 \text{ kg}}$

$h = 190$ cm : nearest : person 3 (191 cm), 4 (182 cm), 6 (178 cm)

$\hat{y}_{KNN} = \frac{100+80+83}{3} = \boxed{87.667 \text{ kg}}$

c) $h = 150\,cm$ $\hat{y}_{kNN} = \dfrac{\left(\frac{1}{150-150}\right)(65) + \left(\frac{1}{168-150}\right)(78) + \left(\frac{1}{171-150}\right)(80)}{\left(\frac{1}{150-150}\right) + \left(\frac{1}{168-150}\right) + \left(\frac{1}{171-150}\right)}$

$= \boxed{\frac{\infty}{\infty}}$ = undefined

$h = 155\,cm$ $\hat{y}_{kNN} = \dfrac{\left(\frac{1}{155-150}\right)(65) + \left(\frac{1}{168-155}\right)(78) + \left(\frac{1}{171-155}\right)(80)}{\left(\frac{1}{155-150}\right) + \left(\frac{1}{168-155}\right) + \left(\frac{1}{171-155}\right)}$

$= \boxed{70.708\ kg}$

$h = 165\,cm$ $\hat{y}_{kNN} = \dfrac{\left(\frac{1}{168-165}\right)(78) + \left(\frac{1}{171-165}\right)(80) + \left(\frac{1}{178-165}\right)(83)}{\left(\frac{1}{168-65}\right) + \left(\frac{1}{171-165}\right) + \left(\frac{1}{178-165}\right)}$

$= \boxed{79.244\ kg}$

$h = 190\,cm$ $\hat{y}_{kNN} = \dfrac{\left(\frac{1}{191-190}\right)(100) + \left(\frac{1}{190-182}\right)(80) + \left(\frac{1}{190-178}\right)(83)}{\left(\frac{1}{191-190}\right) + \left(\frac{1}{190-182}\right) + \left(\frac{1}{190-178}\right)}$

$= \boxed{96.759\ kg}$