HW4

2022-11-19

1 VC dimension of Binary Decision Trees with Fixed Split Points

If there are l leaves then there are l distinct paths in the tree which correspond to splitting the feature space into l subsets. Therefore, if I take one point from each of the subsets, so I have l points, then if I take any tree $f \in \mathcal{F}$ each leaf will contain 1 point and only one point. Then if I label the leaf the exact label of the single point inside it then I will classify all of the l points correctly. By the definition of how $\mathcal{F}$ is constructed, for every permutation of labels of these l points there is a corresponding tree $f \in \mathcal{F}$ such that the labels of the leaves match the labels of the points. Thus $\mathcal{F}$ has a VC Dimension of at least l.

If there are $l+1$ points, then there must exist at least 2 points a, b belonging to the same sub-feature space. Therefore every tree $f \in \mathcal{F}$ will have a leaf that contains both points a, b. However, one permutation of labels for the $l+1$ points is where a and b have opposite signs. However, Since they are in the same leaf they will be classified with the same label and thus one of the labels is incorrect and $\mathcal{F}$ does not shatter $l+1$ points an thus the VC dimension of $\mathcal{F}$ is l.

2 Topic Modeling with EM

2.1

$$\log \prod^{N} \prod^{M} p(w_n, |d_i, \alpha, \beta)^{q(w_n, d_i)} = \log \prod^{N} \prod^{M} p_{ni}^{q(w_n, d_i)} = \log \prod^{N} \prod^{M} \left(\sum_{k=1}^{K} \beta_{kn} \alpha_{ik} \right)^{q(w_n, d_i)}$$

$$= \sum^{N} \sum^{M} q(w_n, d_i) \log \sum_{k=1}^{K} \beta_{kn} \alpha_{ik}$$

2.2

Bayes rule gives us:

$$p(z_k | d_i, w_n, \alpha^{old}, \beta^{old}) = \frac{p(w_n | z_k, d_i, \alpha^{old}, \beta^{old}) p(z_k | d_i, \alpha^{old}, \beta^{old})}{p(w_n | d_i, \alpha^{old}, \beta^{old})}$$

$$p(z_k | d_i, w_n, \alpha^{old}, \beta^{old}) = \frac{\beta_{kn}^{old} \alpha_{ik}^{old}}{\sum_{k=1}^{K} \beta_{kn}^{old} \alpha_{ik}^{old}}$$

2.3

$$\text{log-liklihood} = \sum^{N} \sum^{M} q(w_n, d_i) \log p(w_n, |d_i, \alpha, \beta) = \sum^{N} \sum^{M} q(w_n, d_i) \log \sum_{k=1}^{K} p(w_n, z_k | d_i, \alpha, \beta)$$

$$= \sum^{N} \sum^{M} q(w_n, d_i) \log \sum_{k=1}^{K} \frac{p(z_k | w_n, d_i, \alpha^{old}, \beta^{old}) p(w_n, z_k | d_i, \alpha, \beta)}{p(z_k | w_n, d_i, \alpha^{old}, \beta^{old})} = \sum^{N} \sum^{M} q(w_n, d_i) \log E_z \left(\frac{p(w_n, z_k | d_i, \alpha, \beta)}{p(z_k | w_n, d_i, \alpha^{old}, \beta^{old})} \right)$$

by Jensens inequality $\log E(x) \geq E(\log x)$:

Also we define $\gamma_{ink} := p(z_k|w_n, d_i, \alpha^{old}, \beta^{old})$

$$\geq \sum^{N}\sum^{M} q(w_n, d_i) E_z \log\left(\frac{p(w_n, z_k|d_i, \alpha, \beta)}{p(z_k|w_n, d_i, \alpha^{old}, \beta^{old})}\right) = \sum^{N}\sum^{M} q(w_n, d_i) \sum_{k=1}^{K} p(z_k|w_n, d_i, \alpha^{old}, \beta^{old}) \log\left(\frac{p(w_n, z_k|d_i, \alpha, \beta)}{p(z_k|w_n, d_i, \alpha^{old}, \beta^{old})}\right)$$

$$= \sum^{N}\sum^{M}\sum_{k=1}^{K} q(w_n, d_i)\gamma_{ink} \log\left(\frac{p(w_n, z_k|d_i, \alpha, \beta)}{\gamma_{ink}}\right) = \sum^{N}\sum^{M}\sum_{k=1}^{K} q(w_n, d_i)\gamma_{ink} \log\left(\frac{\alpha_{ik}\beta_{kn}}{\gamma_{ink}}\right) = A(\alpha, \beta)$$

2.4

$$\mathcal{L} = \sum^{N}\sum^{M}\sum^{K} q(w_n, d_i)\gamma_{ink} \log\left(\frac{\alpha_{ik}\beta_{kn}}{\gamma_{ink}}\right) + \lambda_1\left[1 - \sum^{K}\alpha_{ik}\right] + \lambda_2\left[1 - \sum^{K}\beta_{kn}\right]$$

$$\frac{\partial\mathcal{L}}{\partial\alpha_{ik}} = \sum^{N}\sum^{M}\sum^{K} q(w_n, d_i)\gamma_{ink}\frac{1}{\alpha_{ik}\beta_{kn}}\beta_{kn} - \lambda_1 = \frac{1}{\alpha_{ik}}\sum^{N} q(w_n, d_i)\gamma_{ink} - \lambda_1 = 0$$

Therefore:

$$\frac{1}{\alpha_{ik}} = \frac{\lambda_1}{\sum^{N} q(w_n, d_i)\gamma_{ink}} \rightarrow \alpha_{ik} = \frac{\sum^{N} q(w_n, d_i)\gamma_{ink}}{\lambda_1}$$

$$1 = \sum^{K}\alpha_{ik} = \frac{1}{\lambda_1}\sum^{K}\sum^{N} q(w_n, d_i)\gamma_{ink} \quad \Rightarrow \quad \lambda_1 = \sum^{K}\sum^{N} q(w_n, d_i)\gamma_{ink}$$

Therefore:

$$\alpha_{ik} = \frac{\sum^{N} q(w_n, d_i)\gamma_{ink}}{\sum^{K}\sum^{N} q(w_n, d_i)\gamma_{ink}}$$

$$\frac{\partial\mathcal{L}}{\partial\beta_{kn}} = \sum^{N}\sum^{M}\sum^{K} q(w_n, d_i)\gamma_{ink}\frac{1}{\alpha_{ik}\beta_{kn}}\alpha_{ik} - \lambda_2 = \frac{1}{\beta_{kn}}\sum^{M} q(w_n, d_i)\gamma_{ink} - \lambda_2 = 0$$

Therefore:

$$\frac{1}{\beta_{kn}} = \frac{\lambda_2}{\sum^{M} q(w_n, d_i)\gamma_{ink}} \rightarrow \beta_{kn} = \frac{\sum^{M} q(w_n, d_i)\gamma_{ink}}{\lambda_2}$$

$$1 = \sum^{N}\beta_{kn} = \frac{1}{\lambda_2}\sum^{N}\sum^{M} q(w_n, d_i)\gamma_{ink} \quad \Rightarrow \quad \lambda_2 = \sum^{N}\sum^{M} q(w_n, d_i)\gamma_{ink}$$

Therefore:

$$\beta_{kn} = \frac{\sum^{M} q(w_n, d_i)\gamma_{ink}}{\sum^{N}\sum^{M} q(w_n, d_i)\gamma_{ink}}$$

3 Gradient Computations in Neural Networks

3.1

Lets start by take the derivative of a specific weight $W_{1,(j,k)}$ which is a specific weight of the W_1 matrix connecting the jth element in in h_1 with the kth element of x_i. $x_i \in \mathbb{R}^{784}$

This gives us:

$$\frac{d\mathcal{L}_\rangle}{dW_{1,(j,k)}} = \sum_{t=1}^{H} \frac{d\mathcal{L}_i}{df} \frac{df}{dz_3} \frac{dz_3}{dh_{2,t}} \frac{dh_{2,t}}{dz_{2,t}} \frac{dz_{2,t}}{dh_{1,j}} \frac{dh_{1,j}}{dz_{1,j}} \frac{dz_{1,j}}{dW_{1,(j,k)}}$$

$$= \sum_{t=1}^{H} \left[-\frac{y_i}{f(x_i)} + \frac{1-y_i}{1-f(x_i)} \right] \cdot f(x_i)(1-f(x_i)) \cdot W_{3,t} \cdot h_{2,t}(1-h_{2,t}) \cdot W_{2,(t,j)} \cdot h_{1,j}(1-h_{1,j}) \cdot x_{i,k}$$

Then for the weights W_1 and and N observations we have:

$$\frac{d\mathcal{L}}{dW_1} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}_\rangle}{\partial f} \frac{\partial f}{\partial z_3} \frac{\partial z_3}{\partial h_2} \frac{\partial h_2}{\partial z_2} \frac{\partial z_2}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial W_1}$$

Then if we let x be the matrix of all N observations so it has the shape N x 784

$$\frac{d\mathcal{L}}{dW_1} = \frac{\partial \mathcal{L}}{\partial f} \frac{\partial f}{\partial z_3} \frac{\partial z_3}{\partial h_2} \frac{\partial h_2}{\partial z_2} \frac{\partial z_2}{\partial h_1} \frac{\partial h_1}{\partial z_1} \frac{\partial z_1}{\partial W_1}$$

$$= \left[-\frac{y}{f(x)} + \frac{1-y}{1-f(x)} \right] \cdot f(x)(1-f(x)) \cdot W_3 \cdot h_2(1-h_2) \cdot W_2 \cdot h_1(1-h_1) \cdot x$$

3.2

General form

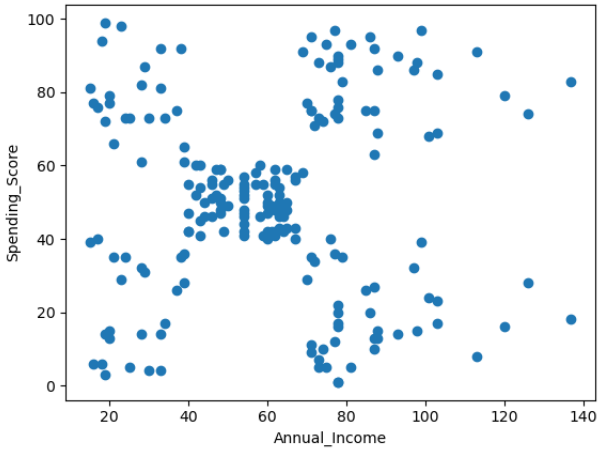
$$\delta_l^i = \frac{\partial \mathcal{L}}{\partial z_l}$$

$$\frac{\partial \mathcal{L}}{\partial z_l} = \delta_{l+1}^i \cdot W_{l+1} \cdot h_l(1-h_l)$$

4 Clustering

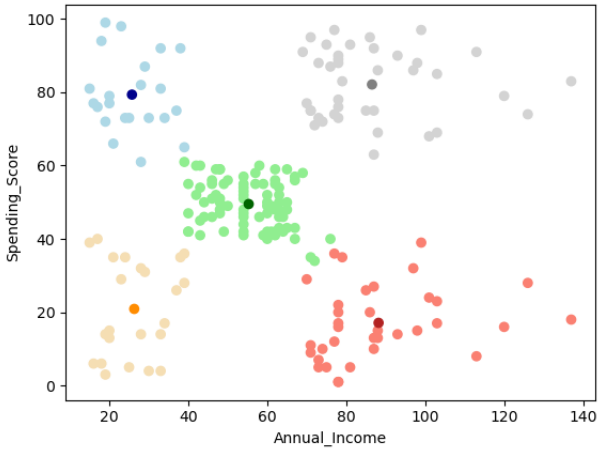
4.1

look at code solutions for more info



4.2

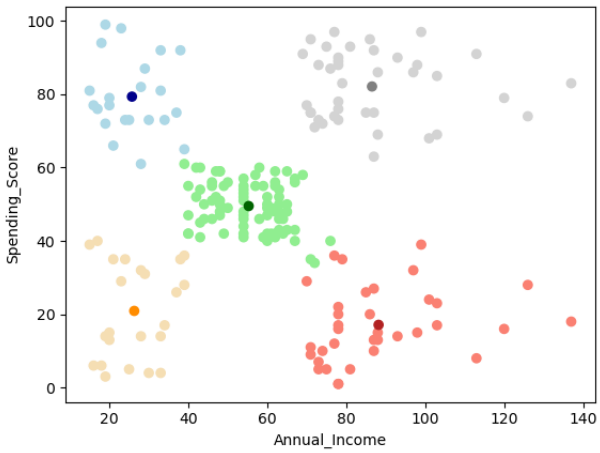
look at code solutions for more info



The clusters at the bottom left, middle, and top right represent customers who spending is proportional to their income while the clusters at the top left and bottom right are customers who overspend and underspend relative tho their incomes.

4.3

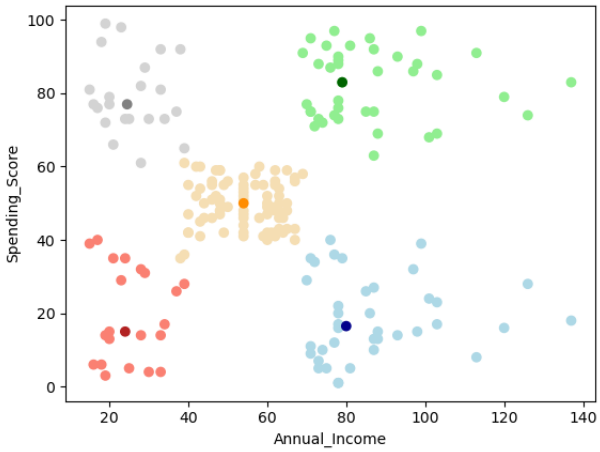
look at code solutions for more info



I get extremely similar but not identical cluster centers between my kmeans implementation vs sklearns implementation. My metric uses euclidean distance to measure the distance from each point to their center and sums the total distances. The total summed distance metric for my implementation had a value of 2602.9725788386436 while the total summed distance metric for sklearns implementation had a value of 2604.025270158349. Since mine is smaller I would say mine performs better.

4.4

look at code solutions for more info



I get similar results but not exactly the same between kmeans and kmedians

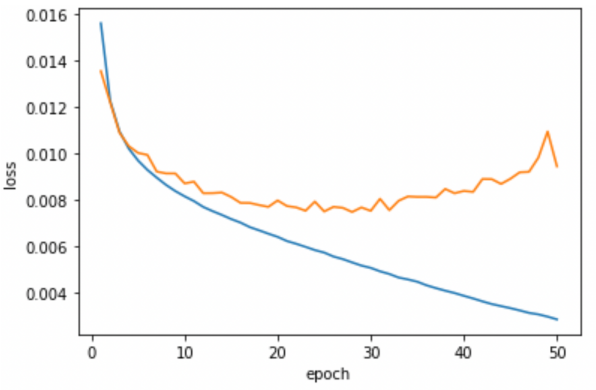
5 Convolutional Neural Network on CIFAR-10

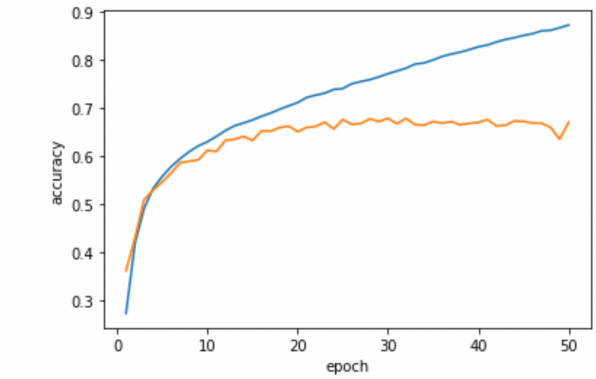
5.1

look at code solutions for more info

5.2

look at code solutions for more info





We see that the model is overfitting since the training accuracy continues to rise while the validation accuracy plateaus at around .65.

5.3

look at code solutions for more info

From the confusion matrix the we see that class 4 and class 6 were the two most misclassified images. This makes sense because class 4 is cats and clas 6 is dogs and these animals look very similar and have very similar features which will result in mislabeling.