AJ-0061-HWY

1. a) $X=\begin{bmatrix}1&2\\3&4\\3&3\\1&1\end{bmatrix}$, $Y=\begin{bmatrix}2\\8\\5\\3\end{bmatrix}$. The MLE estimate for θ using least squares

is $\hat{\theta}=(X^TX)^{-1}X^TY$.

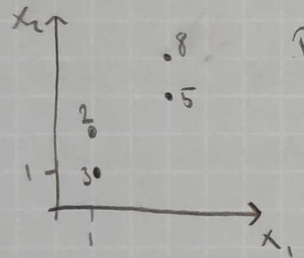
$$X^TX=\begin{bmatrix}30&24\\24&20\end{bmatrix} \Rightarrow (X^TX)^{-1}=\frac{1}{30\cdot 20-24^2}\begin{bmatrix}20&-24\\-24&30\end{bmatrix}=\frac{1}{24}\begin{bmatrix}20&-24\\-24&30\end{bmatrix}=\begin{bmatrix}5/6&-1\\-1&5/4\end{bmatrix}$$

$X^TY=\begin{bmatrix}54\\44\end{bmatrix}$ so $\hat{\theta}=(X^TX)^{-1}X^TY=\begin{bmatrix}5/6&-1\\-1&5/4\end{bmatrix}\begin{bmatrix}54\\44\end{bmatrix}=\begin{bmatrix}45-44\\-54+55\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}$

$x_*=[2\ 3] \Rightarrow \hat{y}_*=x\hat{\theta}=[2\ 3]\begin{bmatrix}1\\1\end{bmatrix}=5$

b) Further away, the point of regularization is that sometimes, using MLE induces overfit so we settle for an estimate that is worse on training data but hopefully does not overfit as much.

c) The cost function in the regression tree is mean squared error.



Possible splits: $x_1=2$, $x_2=1.5$, $x_2=2.5$, $x_2=3.5$

$x_1=2$: $\bar{y}_< = \frac{2+3}{2}=\frac{5}{2}$, $\bar{y}_> = \frac{8+5}{2}=\frac{13}{2}$ $\Rightarrow$ cost $=\left(\frac{1}{2}\left((2-\frac{5}{2})^2+(3-\frac{5}{2})^2\right)+\frac{1}{2}\left((8-\frac{13}{2})^2+(5-\frac{13}{2})^2\right)\right)$

$=\frac{1}{4}+\frac{9}{4}=\frac{5}{2}$

$x_2=1.5$: $\bar{y}_\leq = 3$, $\bar{y}_> = \frac{2+5+8}{3}=\frac{15}{3}=5$ $\Rightarrow$ cost $=\left((3-3)^2+\frac{1}{3}\left((2-5)^2+(5-5)^2+(8-5)^2\right)\right)$

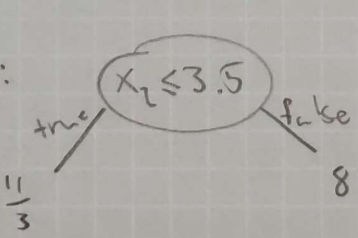
$=0+\frac{2}{3}\cdot 9=6$

$x_2=2.5$: same as $x_1=2$ so cost $=\frac{5}{2}$

$x_2=3.5$: $\bar{y}_\leq=\frac{2+3+5}{3}=\frac{11}{3}$, $\bar{y}_>=8$ $\Rightarrow$ cost $=\left(\frac{1}{3}\left((2-\frac{11}{3})^2+(3-\frac{11}{3})^2+(5-\frac{11}{3})^2\right)+(8-8)^2\right)=$

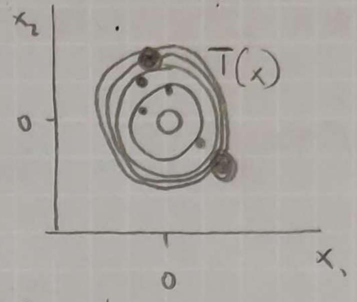
$=\frac{5}{3}+0=\frac{5}{3}<\frac{5}{2}$

So, split at $x_2=3.5$:



$x^*=[2\ 3]$ so $x_2^*=3<3.5$ so $\hat{y}^*=\frac{11}{3}$

1 d) The value of y seems to increase with distance to the origin, and the rate that y increases by seems to increase as well. A suitable transformation of $x = [x_1, x_2]$ could then be $\|x\|^2 = x_1^2 + x_2^2$ since the level curves of that transformation form concentric circles about the origin. Let $T(x) = x_1^2 + x_2^2$



Level curves and example data

AJ-0061-HWY

2. a) For QDA, we assume that the data follows a normal distribution, $X|Y=y \sim N(\mu_y, \sigma_y^2)$ (one dim. data so just normal)

Encode asthma as 1 and no asthma as 0.

Then $X|Y=1 \sim N(81, 81^2)$ and $X|Y=0 \sim N(100, 100^2)$.

In this Bayesian setting, the prior for Y is 10% asthma, 90% no asthma, e.g. $Y \sim Ber(\frac{1}{10})$.

$$P(Y=1|X=300) = \frac{f_{N(81,81^2)}(300)P(Y=1)}{f_{N(81,81^2)}(300)P(Y=1) + f_{N(100,100^2)}(300)P(Y=0)} = \frac{1.27 \cdot 10^{-4} \cdot 0.1}{1.27 \cdot 10^{-4} \cdot 0.1 + 5.40 \cdot 10^{-4} \cdot 0.9} =$$

$= 0.0255$

So, the posterior probability of asthma is 2.55%.

b) The normal distribution is symmetric about the mean, but it is not unreasonable to believe that the actual distribution of the test result is highly unsymmetric. For example, the child in a) could get 300 mm but under the normal assumption, it would be equally likely to get -100 mm which is obviously not the case.

c) $X|Y=y \sim Exp(\lambda_y)$, $\lambda_0 = \frac{1}{100}$, $\lambda_1 = \frac{1}{81}$, $Y \sim Ber(\frac{1}{10})$.

$$P(Y=1|X=x) = \frac{f_{Exp(\frac{1}{81})}(x)P(Y=1)}{f_{Exp(\frac{1}{81})}(x)P(Y=1) + f_{Exp(\frac{1}{100})}(x)P(Y=0)} = \frac{1}{1 + \frac{f_{Exp(\frac{1}{100})}(x)P(Y=0)}{f_{Exp(\frac{1}{81})}(x)P(Y=1)}}$$

$f_{Exp(\frac{1}{100})}(x) = \frac{1}{100}e^{-\frac{x}{100}}$, $f_{Exp(\frac{1}{81})}(x) = \frac{1}{81}e^{-\frac{x}{81}}$, $P(Y=0) = \frac{9}{10}$, $P(Y=1) = \frac{1}{10}$, $81 = 9^2$, so

$$\frac{f_{Exp(\frac{1}{100})}(x)P(Y=0)}{f_{Exp(\frac{1}{81})}(x)P(Y=1)} = \frac{\frac{9}{1000}e^{-\frac{x}{100}}}{\frac{1}{10 \cdot 9^2}e^{-\frac{x}{81}}} = \frac{9^3}{100}e^{\frac{x}{81} - \frac{x}{100}} = \frac{9^3}{100}e^{\frac{19x}{8100}}$$ and we get the result.

d) $P(Y=1|X=50) = \frac{1}{1 + \frac{9^3}{100}e^{\frac{19 \cdot 50}{8100}}} = 0.109$, $P(Y=1|X=300) = \frac{1}{1 + \frac{9^3}{100}e^{\frac{19 \cdot 300}{8100}}} = 0.0636$

Probably not very useful in practice since the posterior probability for having asthma given a very low result (x=50) only came out as ~0.11 compared to the prior probability of 0.1. The test probably has low TPR.

AJ-0061-HWY

Q^M is misclass. error

Q^G is Gini index

3.

a) Split A: $\hat{\pi}_{left,0} = \frac{500}{600}$, $\hat{\pi}_{right,0} = \frac{100}{600}$

$\hat{\pi}_{left,1} = \frac{100}{600}$, $\hat{\pi}_{right,1} = \frac{500}{600}$

$Q^M_{left} = 1 - \frac{500}{600} = \frac{1}{6}$ $Q^M_{right} = 1 - \frac{500}{600} = \frac{1}{6}$

$Q^G_{left} = \frac{5}{6}\left(1-\frac{5}{6}\right) + \frac{1}{6}\left(1-\frac{1}{6}\right)$ $Q^G_{right} = \frac{1}{6}\left(1-\frac{1}{6}\right) + \frac{5}{6}\left(1-\frac{5}{6}\right)$

$= 2 \cdot \frac{5}{36} = \frac{5}{18}$ $= 2 \cdot \frac{5}{36} = \frac{5}{18}$

$cost^M = 600 \cdot \frac{1}{6} + 600 \cdot \frac{1}{6} = 2 \cdot 100 = 200$

$cost^G = 600 \cdot \frac{5}{18} + 600 \cdot \frac{5}{18} = 2 \cdot 100 \cdot \frac{5}{3} = \frac{1000}{3} \approx 333.3$

Split B: $\hat{\pi}_{left,0} = \frac{200}{800}$, $\hat{\pi}_{right,0} = \frac{400}{400}$

$\hat{\pi}_{left,1} = \frac{600}{800}$, $\hat{\pi}_{right,1} = \frac{0}{400}$

$Q^M_{left} = 1 - \frac{600}{800} = \frac{1}{4}$, $Q^M_{right} = 1 - \frac{400}{400} = 0$

$Q^G_{left} = \frac{1}{4}\left(1-\frac{1}{4}\right) + \frac{3}{4}\left(1-\frac{3}{4}\right)$ $Q^G_{right} = 1(1-1) + 0(1-0) = 0$

$= 2 \cdot \frac{3}{16} = \frac{3}{8}$

$Cost^M = 800 \cdot \frac{1}{4} + 400 \cdot 0 = 200$

$cost^G = 800 \cdot \frac{3}{8} + 400 \cdot 0 = 300$

For misclassification error, the two split are equal in terms of loss, but for Gini Index, split B is better.

b) Split B creates a pure node (right branch), which means that we only need to focus on the left branch. This indicates that the Gini index might be better since it values pure nodes higher than misclassification error does.

AJ-0061-HWY

3 c) We can create bootstrap samples by sampling with replacement from the data we have.
We then train a tree classifier on each bootstrap sample and let the final classifier be a majority vote from the classifiers trained on bootstrap samples.

This would decrease the variance since no tree is trained on the full dataset. It does not increase the bias since we combine all classifiers at the end.

d) We would increase the weights on misclassified points and decrease weights for correctly classified points (with respect to majority vote).
In the left branch, we would increase for $Y=1$ and decrease for $Y=0$, and reversed in the right branch.

4. $M=4$, input is $20\times20\times3$, 90-10 train-test split. | AJ-0061-HWY

a) size of $W^{(1)}$ is filter rows $\times$ filter columns $\times$ input channels $\times$ output channels, which is $5\times5\times3\times24$

Size of $b^{(1)}$ is $1\times1\times1\times$ output channels which is $1\times1\times1\times24$

Size of q is $\frac{\text{input rows}}{\text{stride}} \times \frac{\text{input columns}}{\text{stride}} \times$ output channels which is

$10\times10\times24$

b) If "input" refers to q, $W^{(2)}$ has size (input rows $\cdot$ input cols $\cdot$ input chanels) $\times M$ which is $(10\cdot10\cdot24)\times4) = 2400\times4$.

$b^{(2)}$ has size $1\times M$ which is 1×4

z has size $1\times M$ which is 1×4

c) Total number of parameters is $\underbrace{5\cdot5\cdot3\cdot24}_{W^{(1)}} + \underbrace{24}_{b^{(1)}} + \underbrace{2400\cdot4}_{W^{(2)}} + \underbrace{4}_{b^{(2)}} = 11428$

d) In a dense layer, every pair of input i and output j has an individual weight w_{ij}, meaning that every output is h(a linear combination of all the inputs), where h is the activation function.

In a convolutional layer, the same filters are applied to every input (here, the input is a neighborhood of each "pixel") to produce an output. This means that the same weights are used for each input, so the number of parameters is much lower compared to a dense layer with the same number of inputs and outputs.

AJ-0061-HWY

4 e) i) One way to improve the model is to add more layers, e.g. a hidden, dense layer after 4. This would increase training time but hopefully improve error as well.

ii) With the addition of a dense hidden layer, we could also consider dropout; removing some units in the hidden layer each iteration to reduce overfit.

iii) Finally, we could tweak the learning rate of SGD, decreasing it as the accuracy improves during training. This could improve the chances of actually converging to at least a good local minimum of the error function.

AJ-0061-HWY

5a) Males: $TPR = \frac{TP}{TP+FN} = \frac{90}{90+10} = \frac{9}{10} = 0.9$

$FPR = \frac{FP}{FP+TN} = \frac{200}{200+800} = \frac{2}{10} = 0.2$

$Precision = \frac{TP}{TP+FP} = \frac{90}{90+200} = \frac{9}{29} = 0.31$

Females: $TPR = \frac{190}{190+10} = \frac{19}{20} = 0.95$

$FPR = \frac{200}{200+800} = \frac{2}{10} = 0.2$

$Precision = \frac{190}{190+200} = \frac{19}{39} = 0.49$

b) FPR is fair since it is equal for both groups.

TPR is, in my opinion, neither fair nor unfair since it is quite close but not equal. The door working in 90% of cases is still very good even though 95% is better.

Precision is much more different, 31% vs 49% so I would say this is unfair. It is unfair to males, since the algorithm is more likely to let a female non-worker in than a male non-worker. This in itself is not a problem, but the algorithm could probably be tweaked to either be more lenient towards males (increasing Male TPR in the process) or stricter towards females (decreasing Female TPR). Since TPR is already better for females, this would not create other major imbalances.

5 c) The precision is equal for true-positive-rates of 0.0, 0.8 and 1.0. Both 0 and 1 are terrible choices for TPR since they correspond to admitting noone or everyone, respectively. Hence, TPR = 0.8 is the only reasonable choice.
This yield Male FPR of 0.1 and Female FPR of 0.2.

That is a big difference, but since the FPR does not really concern the employees and they have the same probability of having it work, I would say that aspect is fair. However, the TPR decreased for both male and female employees so by getting equal precision, the situation for employees worsened and the point was to improve the situation.
In the end, I believe that the equal precision is not worth the decreased TPR since equality probably should not be achieved by making it worse for everyone.

d) The claim that this model is 95% or 90% accurate for males and females, respectively, is not quite correct. The model has those numbers as true-positive-rates but this is only a measure for how well the model works for the employees, to get the accuracy, you would need to consider how well the model performs on non-employees as well.
That model happens to have the same false-positive-rate for males and females so that is not a huge problem, but I would propose the model yielding 90% true-positive-rate and 10% false-negative rate for both males and females, not only does it perform the same for all employees, it keeps more non-employees out of the building as well (10% FPR compared to 20% for your proposal).