Q1

a) Myself

b) I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

Q5

What is the main downfall of decision trees and how do random forests solve that?

OVERFITTING.

that can be seen in the training accuracy results of a single decision tree vs its validation accuracy.

Q2

$$\hat{w} = \underset{\underline{w}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|y - Xw\|_2^2 + \lambda \|w\|_1 \right\}$$

$$\Rightarrow J_\lambda(w) = \frac{1}{2}\|y\|^2 + \frac{1}{2}\|Xw\|^2 - y^T X w + \lambda \|w\|_1$$

$$= \frac{1}{2}\|y\|^2 + \frac{1}{2} w^T X^T X \underline{w} - y^T X w + \lambda \|w\|_1$$

$$= \frac{1}{2}\|y\|^2 + \frac{1}{2} w^T \Big(\sum_{ij} \underline{X}_i^T \underline{X}_j\Big) \underline{w} - y^T X w + \lambda \|w\|_1$$

$$= \frac{1}{2}\|y\|^2 + \frac{n}{2} w^T w - \underline{y}^T \underline{X} \underline{w} + \lambda \|w\|$$

~~Uncorrelated X_i's~~

$$\cancel{\sim E[X_i X_j^T] - E[X_i]E[X_j] = \text{diag for } X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}}$$

$$\cancel{J_\lambda(\underline{w}) = \frac{1}{2}\|y - \sum_i w_i \underline{X}_i\|^2 + \lambda \|\textstyle\sum w\|_1}$$

$$J_\lambda(w) = \frac{1}{2}\|y\|^2 + \frac{n}{2}\|w\|_2^2 + \lambda\|w\|_1 - y^T \underline{X}\underline{w}$$

Solve All together

$$\frac{\partial}{\partial w}(J_\lambda(w)) = n\underline{w} + \frac{\partial}{\partial \underline{w}} \|w\|_1 - (X^T y) = 0$$

$$= n\,\underline{w} + \mathrm{sig}(\underline{w}) - X^T y = 0$$

$$\underline{w} = \frac{1}{n}\left[X^T y - \mathrm{sign}(\underline{w})(\lambda)\right]$$

Solve for w_i Separately

$$\frac{\partial}{\partial w_i} J_{i,\lambda}(w_i) = \frac{\partial}{\partial w_i}\left[\frac{n}{2} w_i^2 - (y^T \underline{X}_i) w_i + \lambda |w_i|\right]$$

$$= n w_i - y^T \underline{X}_i + \lambda\, \mathrm{sgn}(w_i) = 0$$

$$w_i = \frac{1}{n}\left[y^T X_i - \lambda\, \mathrm{sgn}(w_i)\right]$$

positive

b) $\hat{w}_i = \frac{1}{n}[+y_i^T X_i - \lambda] = +\frac{y_i^T X_i}{n} - \frac{\lambda}{n}$

c) $\hat{w}_i = -\frac{1}{n}[y_i^T X_i + \lambda] = +\frac{y_i^T X_i}{n} + \frac{\lambda}{n}$

- negative

d) if $|y_i^T x_i| < |\lambda|$ then $\hat{w}_i = 0$

e) $\hat{w} = (X^T X + \lambda I)^{-1} X^T y$

$\hat{w}_i = \frac{1}{n+\lambda}(X_i^T y_i)$ $\quad X_i^T y_i = 0$ for $\hat{w}_i = 0$

f) $1e-5 = \lambda$

Q3

d) $z_i \sim N(0, \sigma^2)$ $\quad Pr(|z| \geq t) \leq e^{-t^2/2\sigma^2}$

$Pr(|z_i| \geq 2\sigma\sqrt{\log d}) \leq e^{-(2\sigma\sqrt{\log d})^2/2\sigma^2} = e^{-\frac{4\sigma^2 \log d}{2\sigma^2}}$

$= \frac{1}{d^2}$

$d^*(Pr(|z_i| \geq 2\sigma\sqrt{\log d})) \leq \left(\frac{1}{d^2}\right)\cdot d' \leq 1/d \Rightarrow P(\max|z_i| \geq 2\sigma\sqrt{\log d}) \leq 1/d$

all of them $< 2\sigma\sqrt{\log d}$

e) $w = (X^TX)^{-1}X^T(y) = (X^TX)^{-1}X^T(y^* + z) = w^* + (X^TX)^{-1}Xz$

$= w^* + z'$

$z' \sim (0, \sigma^2 I)$ transformed cov matrix

because $\sigma^2 \, \cancel{(X^TX)^{-1}X^T}X(X^TX)^{-1}I$

I because has orthonormal columns

$\hat{w} = w^* + z'$

$\hat{w}_{top}(s) = T_s(\hat{w}) = T_s(w^* + z')$

f) $e = \hat{w}_{top}(s) - w^*$ (assuming w^* is s-sparse)

Max error occurs when $\hat{w}_{top}(s)$ predicts s-originally 0 entries in w^* to have some value.

$\max(\hat{w}_{top}(s) - w^*)$ will give s-wrongly predicted non-zero entries

2- negative s entries of original weights

$\max(\text{Sparsity}(e)) = 2s$

g) Given $\max|z_i'| \leq 2\sigma\sqrt{\log d}$

$w_{top}(s) = T_s(w^* + z_i)$

Consider max entry

$e = T_s(w^* + z_i) - w^*$

$e_i = w_i^* + z_i' - w_i^*$

$= z_i$ or $-w_i^*$

$T_s(w_i^* + z_i) \begin{cases} |w_i^* + z_i| & \text{if } i \text{ is top } s \\ 0 & \text{otherwise} \end{cases}$

$$e_i = \begin{cases} -z_i & \max|e_i| \leqslant 2\sigma\sqrt{\log d} \\ -W_i^* & \leftarrow \text{Assume} \to 0 \text{ because we pr} \end{cases}$$

(h) $\Pr\left(\|\hat{w}_{top(s)} - w^*\|^2 \leqslant 32\sigma^2 s \log d \right)$

here $\sum_i^n |e_i|^2 \leqslant (4\sigma\sqrt{\log d})^2 \times 2s$ because $|e_i| \neq 0$ only $2s$ times max

$\leqslant (16\sigma^2 \log d) 2s$

$\leqslant 32\sigma^2 \log d$

Given that $\mathcal{E}$

from (d) $\Rightarrow \Pr(\text{not } \mathcal{E}) \leqslant 1/d$

$\Rightarrow \Pr(\mathcal{E}) \geqslant 1 - 1/d$

(i) $\|Xv\|^2 = \|v\|^2$ since X has orthonormal columns

$|\lambda_i| = 1 \ \forall \lambda_i$ in $\mathrm{Spec}(X)$

$\frac{1}{n}\|X(\hat{w}_{top(s)} - w)\| \leqslant \frac{1}{n} 32\sigma^2 s \log d$ with Prob $(1 - 1/d)$

(j) When sparsity comes into play

$\left(\frac{1}{n}\|Xe_{tops}\|^2\right) \leqslant \frac{32\sigma^2 s \log d}{n}$ compared $\frac{\|Xe_{LS}\|}{n} \leqslant \sigma^2 \frac{d}{n}$

with Pr $(1 - 1/d)$

$\Rightarrow$ high probability

$\Rightarrow \frac{1}{n}\|Xe_{tops}\|^2 \leqslant \frac{1}{n}\|Xe_{WLS}\|^2$

$\frac{32\cancel{\sigma^2} s \log d}{\cancel{n}} \leqslant \cancel{\sigma^2} \frac{d}{\cancel{n}}$

when

$$\boxed{s \leqslant \frac{d}{32 \log d}}$$

(4k) The variance would only be σ^2 because that's the variance of the noise

there is a factor of $32 s \log d$ that we pay for not knowing which features to use.

Q4

b) Missing missing features:

the values are filled with the mode of the data feature

Not Numerical Values:

If the string is common enough, we make a one-hot-feature column for it.

Missing Class label:

Remove it.

e) ~~exclamation~~ < 0.5 for all

g) Spam:

some common

money < 0.5
private < 0.5
bank < 0.5

Most Common = 'money'

Titanic:

sex < 0.5
class
Sibsp < 0.5

Most common = Sex

j) when x has very little counts of anything, means we don't have much information about x when

$W > 1e-15$

When label is not spam it is easy to classify because they mostly only have characters and no keywords of the ones above. when

$W < 1e-35$