1. (a)

i. B ii. B, C

(b)

i. CD ii. A iii. C

iv. C v. B vi. B

(c)

i. C,F ii. BD iii. AE

(d)

i. BD ii. BD iii. CE

(e)

A: False. Logistic Regression cannot converge when data is linearly separable.

B: False. more leaves means more complicated hypothesis which will increase variance but decrease bias.

C:

D: True It dependends on support vectors which is a small subset of data.

E:

F: True All the linear + identity activation layers can just be replaced with one linear function $a = W^T x + b$, applied on $\hat{y} = \text{sigmoid}(w^T x + b)$ which is simply a linear classifier.

G: False. Kernelized Perception doesn't maximize the margin, it only separates the data with a boundary.

H: True. $p(x;\theta) = \frac{\pi}{\theta}\left(\frac{x}{\theta}\right)^{\pi-1} \exp\left\{-\left(\frac{x}{\theta}\right)^{\pi}\right\}$

$= \pi x^{\pi-1} \theta^{-\pi} \exp\left(-\theta^{-\pi} x^{\pi}\right)$

$= \pi x^{\pi-1} \exp\left(-\theta^{-\pi} x^{\pi} - \pi \log \theta\right)$,

$b(y) : \pi x^{\pi-1}$

$\eta : -\theta^{-\pi}$

$a(\eta) : \pi \log(\theta)$

$T(y) : x^{\pi}$

(f):

A. YES B. NO C. NO D. YES
E. NO F. YES G. YES H. NO
I. YES J. NO

2. CNN

(a): i, $m = n-d+1$

ii, W is a $(n-d+1)\times n$ matrix in the form:

$$W = \begin{bmatrix} \theta_1\ \theta_2 \cdots \theta_d & & & \\ & \theta_1\ \theta_2 \cdots \theta_d & & \\ & & \theta_1\ \theta_2 \cdots \theta_d & \\ & & \ddots & \\ & & & \theta_1\ \theta_2 \cdots \theta_d \end{bmatrix}$$

each row is a vector $(\theta_1, \theta_2 \cdots \theta_d)$ shifted right by (row-1) position.

iii. only d parameters needed in the case without intercept, its fully connected equivalent on the othe hand will need $(n-d+1)\cdot n$ parameters, ~~[illegible]~~. Computationally since $d \ll n$, a large portion of W will be 0 which means forward/backword propogation will be much faster than a fully connect network. for both f/b propogation, CNN will only take $(n-d+1)\cdot d / (n-d+1)\cdot n = d/n$ of what a fully connect network needs in ~~comput~~ computation.

(b)

i: $\theta^{[1]}$ is a d dimensional vector, it is used to make the diagonal matrix as indicated in (a). i;
$W^{[2]}$ is $(n-d+1)$ dimensional vector;
$b^{[2]}$ is just a scalar.

ii. The parameters of the network are: $\theta^{[1]}$, $W^{[2]}$ and $b^{[2]}$.
$\theta^{[1]} \in \mathbb{R}^{d}$; we CNN in the setting, $z^{[1]}$, $a^{[1]}$ are $\mathbb{R}^{(n-d+1)}$;
$W^{[2]} \in \mathbb{R}^{n-d+1}$, $b^{(2)} \in \mathbb{R}$; hence, $z^{[2]}$, $a^{[2]}$ are $\mathbb{R}$;
First write all the equations in the network. for ease of description, we $W^{[1]}$ as the CNN matrix composed by $\theta^{[1]}$;

Next page →

$W^{[1]}$ is just $\mathbb{R}^{(n-d+1)\times n}$ matrix : $W^{[1]} = \begin{bmatrix} -\theta^{[1]}- & & \\ & -\theta^{[1]}- & \\ & \ddots & \\ & & -\theta^{[1]}- \end{bmatrix}$

$$L(\hat{y}, y) = -\left((1-y)\log(1-\hat{y}) + y\log(\hat{y})\right)$$

$$\hat{y} = a^{[2]} = \sigma(z^{[2]})$$

$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[1]} = \mathrm{Relu}(z^{[1]})$$

$$z^{[1]} = W^{[1]} x \ ;$$

$(W^{[2]})$

$$\frac{\partial L}{\partial W^{[2]}} = -\frac{\partial}{\partial W^{[2]}}\left((1-y)\log(1-\hat{y}) + y\log\hat{y}\right)$$

$$= \left((1-y)/(1-\hat{y}) - y/\hat{y}\right) \cdot \frac{\partial \sigma}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W^{[2]}}$$

$$= \left((1-y)/(1-\hat{y}) - y/\hat{y}\right) \cdot \hat{y}(1-\hat{y}) \cdot a^{[1]}$$

$$= \left(\hat{y}(1-y) - y(1-\hat{y})\right) a^{[1]}$$

$$= (a^{[2]} - y) a^{[1]} \ ; \quad W^{[2]}, a^{[1]} \in \mathbb{R}^{n-d+1} \ ; \quad a^{[2]} \in \mathbb{R} \ ;$$

$(b^{[2]})$ =

$$\frac{\partial L}{\partial b^{[2]}} = -\frac{\partial}{\partial b^{[2]}}\left((1-y)\log(1-\hat{y}) + y\log\hat{y}\right)$$

$$= \left((1-y)/(1-\hat{y}) + y/\hat{y}\right) \cdot \frac{\partial \sigma}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}}$$

$= a^{[2]} - y \ ;$ $b^{[2]}, a^{[2]}$ both are scalars ;

$(\theta^{[1]})$ = $\Rightarrow$ NEXT PAGE

$(\theta^{[1]})$:

$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial a^{[1]}} \frac{\partial a^{[1]}}{\partial Z^{[1]}} \frac{\partial Z^{[1]}}{\partial W^{[1]}}$$

$$= (a^{[2]} - y) \cdot W^{[2]} \cdot \beta \circ X,$$ "$\circ$" here is outer product.

where $\beta = \frac{\partial a^{[1]}}{\partial Z^{[1]}}$ is just element wise derivative of $a^{[1]}_i$ w.r.t $Z^{[1]}_i$, ~~wth~~ with the ith element 1 when $Z^{[1]}_i > 0$, otherwise 0;

The dimension of $\frac{\partial L}{\partial W^{[1]}}$:

in the ~~above~~ expression $(a^{[2]} - y) \cdot W^{[2]} \cdot \beta \circ X$, $(a^{[2]} - y)$ is a scalar; $W^{[2]}$ is $\mathbb{R}^{n-d+1}$; β is $\mathbb{R}^{n-d+1}$; X is $\mathbb{R}^n$; ~~wh~~ we have $\frac{\partial L}{\partial W^{[1]}}$ is $\mathbb{R}^{(n-d+1)\times n}$, the same as $W^{[1]}$;

For simplicity, let $A = \frac{\partial L}{\partial W^{[1]}}$; the gradient w.r.t $\theta^{[1]}$ is just an ~~n-d+~~ $\mathbb{R}^d$ vector, it's ith element is given by:

$$\left(\frac{\partial L}{\partial \theta^{[1]}}\right)_i = \sum_{j=1}^{n-d+1} A_{(j,i)}; \quad 1 \le i \le d;$$

Update rule is:

$W^{[2]} := W^{[2]} - \alpha \frac{\partial L}{\partial W^{[2]}}$, $\frac{\partial L}{\partial W^{[2]}} = (a^{[2]} - y)\, a^{[1]}$

$b^{[2]} := b^{[2]} - \alpha \frac{\partial L}{\partial b^{[2]}}$, $\frac{\partial L}{\partial b^{[2]}} = a^{[2]} - y$

$\theta^{[1]} := \theta^{[1]} - \alpha \frac{\partial L}{\partial \theta^{[1]}}$, $\left(\frac{\partial L}{\partial \theta^{[1]}}\right)_i = \sum_{j=1}^{n-d+1} A_{(j,i)}$, $A_{(j,i)}$ denotes the jth row, ith column.

(c): Now for $L(\hat{y},y)=\frac{1}{2}(\hat{y}-y)^2$: use chain rule of derivation

$$\frac{\partial L}{\partial W^{[2]}}=(\hat{y}-y)\cdot\hat{y}(1-\hat{y})\cdot a^{[1]}$$
$$=(a^{[2]}-y)\cdot a^{[2]}(1-a^{[2]})\cdot a^{[1]};$$

dimension of $\frac{\partial L}{\partial W^{[2]}}\in\mathbb{R}^{n-d-1}$;

$$\frac{\partial L}{\partial b^{[2]}}=(\hat{y}-y)\cdot\hat{y}(1-\hat{y})$$
$$=(a^{[2]}-y)a^{[2]}(1-a^{[2]});$$

$\frac{\partial L}{\partial b^{[2]}}\in\mathbb{R}$;

$$\frac{\partial L}{\partial W^{[1]}}=\frac{\partial L}{\partial a^{[2]}}\circ\frac{\partial a^{[2]}}{\partial z^{[2]}}\cdot\frac{\partial z^{[2]}}{\partial a^{[1]}}\frac{\partial a^{[1]}}{\partial z^{[1]}}\frac{\partial z^{[1]}}{\partial W^{[1]}}$$
$$\cancel{(a^{[2]}-y)\,a^{[2]}(1-a^{[2]})\,W^{[2]}\circ X;\ \text{when } z^{[1]}\geq 0}$$
$$\cancel{0};\ \cancel{\text{when } z^{[1]}<0};$$
$$=(a^{[2]}-y)a^{[2]}(1-a^{[2]})\cdot W^{[2]}\cdot\beta\circ X;$$

the above $\beta=\frac{\partial a^{[1]}}{\partial z^{[1]}}$ is just a $\mathbb{R}^{n-d+1}$ vector, with the ith element 1 if $z_i^{[1]}\geq 0$, otherwise $\beta_i=0$; the last product "$\circ$" is an outer product.

$\frac{\partial L}{\partial W^{[1]}}\in\mathbb{R}^{(n-d+1)\times n}$; for simplicity, let $\frac{\partial L}{\partial W^{[1]}}=A$;

the ith element of $\theta^{[1]}$'s gradient is:

$$\left(\frac{\partial L}{\partial\theta^{[1]}}\right)_i=\sum_{j=1}^{n-d+1}A_{(j,i)}.\quad 1\leq j\leq n-d+1;\quad 1\leq i\leq d,$$ $A_{(j,i)}$ is the jth row and ith column.

update rules:

$W^{[2]}:=W^{[2]}-\alpha\frac{\partial L}{\partial W^{[2]}}$; $\frac{\partial L}{\partial W^{[2]}}=(a^{[2]}-y)\cdot a^{[2]}(1-a^{[2]})\cdot a^{[1]}$;

$b^{[2]}:=b^{[2]}-\alpha\frac{\partial L}{\partial b^{[2]}}$; $\frac{\partial L}{\partial b^{[2]}}=(a^{[2]}-y)\,a^{[2]}(1-a^{[2]})$

$\theta^{[1]}:=\theta^{[1]}-\alpha\frac{\partial L}{\partial\theta^{[1]}}$; $\left(\frac{\partial L}{\partial\theta^{[1]}}\right)_i=\sum_{j=1}^{n-d+1}A_{(j,i)}$.

$A=\frac{\partial L}{\partial\theta^{[1]}}$ as described above; $A_{(j,i)}$ denotes the jth row, ith Column of matrix A;

3. Linearity of Multinomial Naive Bayes.

Naive Bayes model contains $2k+1$ parameters, k is the size of vocab.

$y \sim \text{Bernoulli}(\theta)$,

$x_i | y=0 \sim \text{Bernoulli}(\theta_{i|y=0})$

$x_i | y=1 \sim \text{Bernoulli}(\theta_{i|y=1})$

All θ, $\theta_{i|y=1}$ and $\theta_{i|y=0}$ are given in lecture notes.

Naive Bayes decision boundary is given by:

$p(y|x) = 0.5$, from lecture notes, we know that:

$$\cancel{= p(x|y=1)\, p(y=1) / (p(x|y=0)\, p(y=0) + p(x|y=1)\, p(y=1))}$$

$$\cancel{\Rightarrow p(x|y=1)\, p(y=1) = p(x|y=0)\, p(y=0)}$$

$$\cancel{\Rightarrow \theta_{x|y=1}\, \theta = \theta_{x|y=0}(1-\theta) \Rightarrow \frac{\theta_{x|y=1}}{\theta_{x|y=0}} = \frac{1-\theta}{\theta}}$$

~~take log on both side, we have:~~

$$\cancel{\log \frac{\theta_{x|y=1}}{\theta_{x|y=0}} + \log \frac{\theta}{1-\theta} = 0}$$

$p(y=1|x)$

$$= p(x|y=1)\, p(y=1) / p(x)$$

$$= (\Pi_{i=1}^{n} (p(x_i|y=1)))\, p(y=1) / ((\Pi_{i=1}^{n} p(x_i|y=0))\, p(y=0) + (\Pi_{i=1}^{n} p(x_i|y=1))\, p(y=1))$$

Set it to 0.5, we have:

$$(\Pi_{i=1}^{n} p(x_i|y=1))\, p(y=1) = (\Pi_{i=1}^{n} p(x_i|y=0))\, p(y=0)$$

Which implies: $\dfrac{\Pi_{i=1}^{n} p(x_i|y=1)}{\Pi_{i=1}^{n} p(x_i|y=0)} = \dfrac{p(y=0)}{p(y=1)}$

take log on both sides and move the right to left:

$$\sum_{i=1}^{n} \log \frac{p(x_i|y=1)}{p(x_i|y=0)} + \log \frac{p(y=1)}{p(y=0)} = 0,$$

the duplicated words can be just replaced by $\phi(x_i) \cdot \log \dfrac{p(x_i|y=1)}{p(x_i|y=0)}$,

So, the above equation can be written as:

$$\phi(x)^T \cdot \log \frac{\theta_{x|y=1}}{\theta_{x|y=0}} + \log \frac{\theta}{1-\theta} = 0,$$

With: $m = \log \dfrac{\theta_{x|y=1}}{\theta_{x|y=0}}$ is a k-dimensional vector;

$c = \log \dfrac{\theta}{1-\theta}$ is a scalar.

4. Kernels.

(a). i. $|\Sigma|^n$

ii. λ^4

iii. Suppose $\psi(x) = \phi(x)/\sqrt{K_{ser}(x,x)}$;

We have $K_{norm}(x,z) = \psi(x)^T\psi(z) = \cancel{K_{ser}(x,z)/\sqrt{K_{ser}(x,x)K_{ser}}}$

$= K_{ser}(x,z)/\sqrt{K_{ser}(x,x)K_{ser}(z,z)}$

From class we know that if a matrix can be written as inner product ~~of~~ of two features $\psi(x)$, $\psi(z)$, it is a kernel.

iv. normized kernel can avoid a feature component to be scaled very small for long non-contiguous substrings. normalization makes the string kernel relatively invariant to the length of the document.

v. $2\lambda^4+\lambda^6$

(b). Refer to the problem set solution, we can know that, if K is a kernel, It's polynomial $p(K(x,z))$ is still a kernel.

i. $K_1(x,z) = \exp(K(x,z))$

$= \sum_{i=0}^{\infty}\frac{1}{i!}(K(x,z))^i$ is a polymial of a kernel, so it is still a kernel.

ii. $K_2(x,z) = \exp\left(-\frac{\|x-z\|^2}{\delta^2}\right)$

$= \frac{1}{\exp\left(\frac{\|x\|^2+\|z\|^2}{\delta^2}\right)}\exp\left(\frac{2x^Tz}{\delta^2}\right)$

~~Let $G_1(\|x\|^2+\|z\|^2)$~~ ~~$G_1 = 2\exp(\|x\|^2+\|z\|^2)$, matrix~~

$= \frac{1}{\exp\left(\frac{\|x\|^2}{\delta^2}\right)}\cdot\frac{1}{\exp\left(\frac{\|z\|^2}{\delta^2}\right)}\cdot\exp\left(\frac{2x^Tz}{\delta^2}\right)$.

Let ~~$G_1(x,z)$~~

$G_1(x,z) = \psi(x)\psi(z)$ where $\psi(x) = \frac{1}{\exp(\|x\|^2/\delta^2)}$ a scalar,

~~as~~ we know that $G_1(x,z)$ is a kernel;

Let $G_2(x,z) = \exp\left(\frac{2x^Tz}{\delta^2}\right) = \exp(k_2(x,z))$ where $k_2(x,z) = 2x^Tz/\delta^2$ is a kernel, from i we know G_2 is also a kernel.

$K_2(x,z) = G_1(x,z)\,G_2(x,z)$, as proved in ps 4.e, $K_2(x,z)$ is also a kernel.

iii. From Gaussian integral we have:

$$\sqrt{\pi/a} = \int_{-\infty}^{\infty} \exp(-a(t-b)^2)\,dt,$$

which implies:

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi/a}} \exp(-a(t-b)^2)\,dt \;;$$

together with:

$$K(x,z) = \exp\left(\tfrac{-1}{2}(x-z)^2\right), \quad \text{set } a = \tfrac{1}{2},\ b = (x+z) \;:$$

$$K(x,z) = \exp\left(-\tfrac{1}{2}(x-z)^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(t+x+z)^2\right) dt$$

$$= \int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}\left(x^2+z^2-2xz\right)\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(t^2+x^2+z^2+2xt+2zt+2xz\right)\right) dt$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\left(t^2+2x^2+2z^2+2xt+2zt\right)\right) dt$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{4}\left((t+2x)^2+(t+2z)^2\right)\right) dt$$

$$= \int_{-\infty}^{\infty} (2\pi)^{-\frac{1}{4}} \exp\left(-\left(\frac{t+2x}{2}\right)^2\right) \cdot (2\pi)^{-\frac{1}{4}} \exp\left(-\left(\frac{t+2z}{2}\right)^2\right) dt$$

This gives us the final feature mapping function:

$$\phi(x,t) = (2\pi)^{-\frac{1}{4}} \exp\left(-\left(\frac{t+2x}{2}\right)^2\right)$$

(c): Kernelizing K-means.

The i-class centroid is now an infinite dimension parameter which cannot be represented by computer, updating centroid directly is changed to update the indentities in Centroids calculation.

~~u_j~~

After Kernelizing, u_j is still a linear combination of $\phi(x^{(i)})$:

$$u_j = \sum_{i=1}^{m} r_i \phi(x^{(i)}), \quad \text{where } r_i = \frac{1}{N_j} \text{ for } C^{(i)} = j, \text{ otherwise } r_i = 0.$$

The norm is now:

$$\|\phi(x^{(i)}) - u_j\|^2 = \left(\phi(x^{(i)})^T \phi(x^{(i)}) - 2\phi(x^{(i)})^T u_j + u_j^T u_j\right)$$

Plug in linear combination of u_j:

$$\begin{aligned}
\|\phi(x^{(i)}) - u_j\|^2 &= K(x^{(i)}, x^{(i)}) - 2\phi(x^{(i)})^T \sum_{j=1}^{m} r_j \phi(x^{(j)}) \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{m} r_i r_j \phi(x^{(i)})^T \phi(x^{(j)}) \\
&= K(x^{(i)}, x^{(i)}) - \sum_{j=1}^{m} 2 r_j K(x^{(i)}, x^{(j)}) \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{m} r_i r_j K(x^{(i)}, x^{(j)})
\end{aligned}$$

So the update rule should be changed to:

$$\tilde{c}^{(i)} = \arg\min_j \left(K(x^{(i)}, x^{(i)}) - 2\sum_{k=1}^{m} 2 r_k K(x^{(i)}, x^{(k)}) + \sum_{k=1}^{m}\sum_{l=1}^{m} r_k r_l K(x^{(k)}, x^{(l)}) \right),$$

5. Trees and Random Forests.

(a). ~~[illegible]~~ .

i. First prove that Gini loss is strictly concave:

For the 2 class case,

$G(R_m) = p_{m1}(1-p_{m1}) + p_{m2}(1-p_{m2})$; since $p_{m1}+p_{m2}=1$, we have:

$G(R_m) = 2p_{m1}(1-p_{m1}) = 2p_{m1} - 2p_{m1}^2$;

take the 2nd order derivative of G w.r.t R_m:

$G'' = -2 < 0$, which means Gini is strictly concave.

Then Let's prove that the weighted Gini loss is less or equal than the parent Gini. for simplicity, p is used for p_{m1} for parent Gini:

The parent Gini loss: $G(R) = 2p(1-p)$.

The weight Gini loss of the Children:

$$G(R_1, R_2) = \min_{j,t} \frac{|R_1|}{|R_1|+|R_2|} L(R_1) + \frac{|R_2|}{|R_1|+|R_2|} L(R_2)$$

$$= \min_{j,t} q \cdot 2 p_{11}(1-p_{11}) + (1-q) \cdot 2 \cdot p_{21}(1-p_{21})$$

$$\leq 2\left(q\, p_{11}(1-p_{11}) + (1-q)\, p_{21}(1-p_{21})\right)$$

p_{11} means the first child's ~~proportion of~~ class 1 samples proportion;

p_{21} means the second child's class 2 examples proportion;

p means the class 1 samples proportion of the whole dataset;

~~[illegible]~~ We have $p_{11}|R_1| + p_{21}|R_2| = p(|R_1|+|R_2|)$, R_1, R_2 denotes the first/second trees. it is obvious that ~~p_{11}, p_{21} [illegible]~~ p lies in between p_{11} and p_{21}.

According to Jensen's inequality and strict concavity, we have

$$G(R_1, R_2) \leq 2q\, p_{11}(1-p_{11}) + 2(1-q)\, p_{21}(1-p_{21})$$

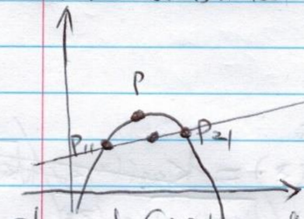
$$\leq 2p(1-p)$$

$$= G(R),$$

Which is what need to be proved in i.

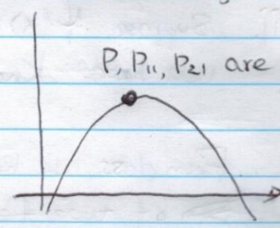
ii. in above explaination, we know that p lies in between p_{11} and p_{21}. if $p = p_{11} = p_{21}$ (in 2 class case, can be generalized to k classes), which means the split doesn't change the class proportions in respective nodes, then Gini Loss would not change after the split. in the stricty concave setting, because after split, class proportions of sub nodes doesn't change. Next →

It can be illustrated in the diagram,: Left one is a Gini changed; Right one is the case that Gini doesn't change.



Changed Gini loss after split.
the segment $P_{11}P_{21}$ lies below the curve of Gini function

P, P_{11}, P_{21} are the same,



in this case the segment $P_{11}P_{21}$ is just a same point as P
Gini loss doesn't change after split.

iii. in the misclassification case, apart from having all class proportions to be the same, there is more cases where we could have the Gini loss unchanged after split.
Suppose the max proportion of a class are all the same in ~~sub trees and~~ each of the children and in parent, the Gini loss will keep the same after the split.

Suppose $M(R) = 1 - \max_k p_k$ is the parent loss;
$M(R_1) = 1 - \max_k p_{1k}$ is the first child loss;
$M(R_2) = 1 - \max_k p_{2k}$ is the second child loss
as long as $\max_k p_k = \max_k p_{1k} = \max_k p_{2k}$,

the weighted Gini loss will keep the same as parent.
$M(R) = q(1 - \max_k p_{1k}) + (1-q)(1 - \max_k p_{2k}) = M(R_1, R_2)$.

(b) : Random Forests.

i. Prove that:

$$Var(\hat{T}) = \rho\delta^2 + \frac{1-\rho}{B}\delta^2.$$

Use the fact that:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y);$$

We have:

$$Var(\hat{T}) = Var\left(\frac{1}{B}\sum_{i=1}^{B} T_i(X)\right)$$

$$= \frac{1}{B^2} Var\left(\sum_{i=1}^{B} T_i\right)$$

$$= \frac{1}{B^2}\left(\sum_{i=1}^{B} Var(T_i) + 2\sum_{j=2}^{B}\sum_{i<j} cov(T_i, T_j)\right)$$

$$= \frac{1}{B^2}\left(B\delta^2 + B(B-1)\cdot\rho\cdot\delta^2\right)$$ [Choose 2 from B samples]

$$= \delta^2\left(\frac{1}{B} + \frac{B-1}{B}\rho\right)$$

$$= \delta^2\left(\rho + \frac{1-\rho}{B}\right)$$

$$= \rho\delta^2 + \frac{1-\rho}{B}\delta^2,$$ proved.

ii. Since $Var(\hat{T}) = \rho\delta^2 + \frac{1-\rho}{B}\delta^2$, for the case of random forest, it will have de-correlation effect, which means value ρ above will be small, which means $Var(\hat{T})$ in random forest is going to be reduced. This makes bagging have higher variance.