SECTION A – True / False

1. Bagging reduces the chance of overfitting, making the model more adaptable to unseen data. TRUE
2. Averaging predictions reduces fluctuations in data. TRUE
3. In boosting, the weights of data points change at every step to make the next gradient descent more accurate. FALSE
4. Sampling without randomness can introduce sampling bias. TRUE
5. If a model is overfit, training error can be 0% while testing error can be 100%. TRUE
6. Regularization simplifies the model, which decreases bias. FALSE (bias increases & Varience decreases)
7. Increasing the depth of a decision tree always prevents overfitting. FALSE
8. Every classifier makes assumptions about the data. TRUE
9. Random Forests are more accurate than single decision trees because they combine bagged trees. TRUE
10. Irreducible error is the lower bound on error due to inherent noise in the data. TRUE

SECTION B – ERM and SVM

1. Fill the following table:

Model	Loss Function	Regularizer
SVM	Hinge loss $\max(0, 1-y\hat{y})$	L2 norm $\|w\|^2$
LASSO	square loss $(y-\hat{y})^2$	L1 norm $\sum\|w_i\|$
RIDGE	square loss	L1 norm

3. Short Answer:

(a) Which loss functions can be optimized using gradient descent? Why?

(b) Which loss functions can be optimized using Newton's method? Why?

a) → all the differentiable loss functions like, square loss, exp loss, hinge loss can be optimized using gradient descent as we have to calculate gradient of the loss function while optimizing it.

b) → Loss function must be twice differentiable ex., square loss

SECTION C – Bias and Variance

1. Explain one major reason why underfitting occurs.
2. If both training and test error remain high, what does this imply about the data?
3. Explain how bagging reduces variance.
4. Explain the effect of boosting on bias and variance.

1) → Underfitting occurs when the model is too simple & it can't capture the pattern in data, one of the main reason of underfitting is having high bias which usually occurs when using strong regularization or using less complex model (like LR model for highly non-linear data) or when important features are not included.

2) → it represents underfitting ie., the model was unable to learn properly resulting in high bias. The relation betn input & output features is very complex or not at all relatable

3) → Bagging reduces variance by training multiple models on different randomly sampled versions of training data & then averaging their predictions as each model sees different samples of data, their individual errors are uncorrelated & hence by averaging out these errors might get cancelled leading to less variance model overall

4) → Boosting trains models sequentially, where each new model focuses more on the wrong predictions made earlier, due to this the bias decreases significantly & the accuracy increases. Incase of variance, boosting may decrease it initially but after many iterations, overfitting may occur as sequential implementation of models makes the process more noise sensitive & hence variance may starts to rise.

SECTION E – Decision Trees

1. Derive that the optimal prediction at a leaf (with squared loss) is the mean.
2. What are the max/min Gini impurity values for 3 classes?
3. Why are decision trees myopic learners?
4. Explain two methods to avoid overfitting in decision trees.

1) → let a leaf contains n samples with target values: $y_1, y_2, \dots y_n$

& the prediction be c

∴ square loss $L(c) = \sum_{i=1}^{n} (y_i - c)^2$

finding c st. L is min

$$\therefore \frac{\partial}{\partial c} L(c) = -\sum_{i=1}^{n} 2(y_i - c) = 0$$

$$\therefore c = \frac{\sum y_i}{n}$$ i.e., mean of target values in the leaf

2) → $\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$ (p_i is i^{th} class probability)

for 3 class, $n = 3$ → $\sum p_i = 1$

$\text{Gini}|_{min}$ is when any one p_i is 1 & rest all are 0

$$\therefore \text{Gini}|_{min} = 1 - (1^2 + 0^2 + 1^2) = 0$$

$\text{Gini}|_{max}$ when are classes are equally likely

$$\therefore \text{Gini}|_{max} = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = \frac{2}{3}$$

3) → while building a tree, at each node algorithm evaluates all possible splits & chooses the split which will cause maximum immediate reduction in impurity ignoring all the future possibilities

a slightly wrong split at the moment may lead to better tree overall but decision tree ignores that aspect thats why it is also known as

myopic or greedy learner

4) → ① as the tree grows continuously it may try to fit the noise resulting in overfitting, hence pre-pruning / early stopping stops the tree from from growing before it perfectly fits the training data to avoid overfitting for this we can i) set max depth ii) set minimum samples per node

② post-pruning allows the tree to grow fully & then removes branches that do not improve the performance on validation data.

SECTION F – Boosting & Bagging

1. Can Random Forests use the same data for training and testing? Justify.
2. Explain the key difference between bagging and boosting.

1) → we should never use same data for training & testing for any model, because doing so can give very excellent result on testing set but will perform poorly on unseen data. As it causes memorization of already seen data from training

2) →

Bagging	Boosting
1) Multiple models are trained independently on different sample data.	1) Multiple models are trained sequentially on same training dataset
2) Final prediction is average of the predictions of all the models	2) final prediction is weighted combination of all models
3) Works best for high variance models as decision trees (as it ↓ variance)	3) Converts weak learners to strong learners (bias decreases)