

[3 pts] What strategies can help reduce overfitting in decision trees?

- ☒ Pruning
- ☐ Make sure each leaf node is one pure class
- ☒ Enforce a minimum number of samples in leaf nodes
- ☒ Enforce a maximum depth for the tree

[3 pts] Neural networks

- ☐ optimize a convex cost function
- ☐ always output values between 0 and 1
- ☒ can be used for regression as well as classification
- ☒ can be used in an ensemble

[3 pts] Which of the following are true about generative models?

- ☒ They model the joint distribution $P(\text{class} = C \text{ AND sample} = \mathbf{x})$
- ☐ The perceptron is a generative model
- ☒ They can be used for classification
- ☒ Linear discriminant analysis is a generative model

[3 pts] Which of the following methods can achieve zero training error on *any* linearly separable dataset?

- ☒ Decision tree
- ☐ 15-nearest neighbors
- ☒ Hard-margin SVM
- ☒ Perceptron

[3 pts] The kernel trick

- ☐ can be applied to every classification algorithm
- ☐ is commonly used for dimensionality reduction
- ☐ changes ridge regression so we solve a $d \times d$ linear system instead of an $n \times n$ system, given n sample points with d features
- ☒ exploits the fact that in many learning algorithms, the weights can be written as a linear combination of input points

[3 pts] Suppose we train a hard-margin linear SVM on $n > 100$ data points in \mathbb{R}^2 , yielding a hyperplane with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the maximum possible number of support vectors for the new hyperplane (assuming the $n + 1$ points are linearly separable)?

- ☐ 2
- ☐ n
- ☐ 3
- ☒ $n + 1$

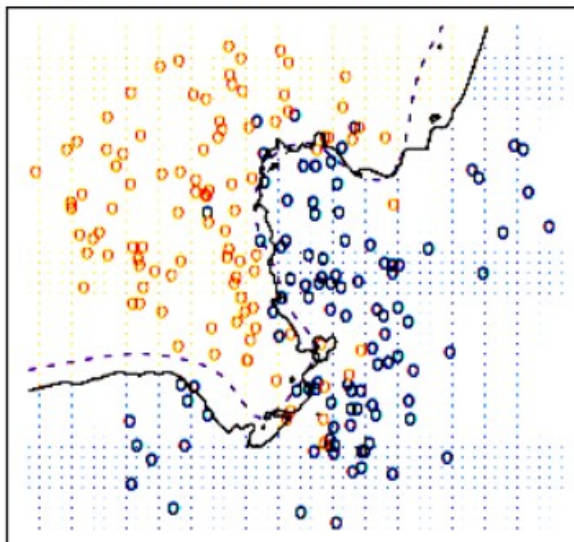
[3 pts] In neural networks, nonlinear activation functions such as sigmoid, tanh, and ReLU

- ☐ speed up the gradient calculation in backpropagation, as compared to linear units
- ☒ help to learn nonlinear decision boundaries
- ☐ are applied only to the output units
- ☐ always output values between 0 and 1

[3 pts] Which of the following can help to reduce overfitting in an SVM classifier?

- ☒ Use of slack variables
- ☐ High-degree polynomial features
- ☐ Normalizing the data
- ☐ Setting a very low learning rate

[3 pts] Which value of k in the k -nearest neighbors algorithm generates the solid decision boundary depicted here? There are only 2 classes. (Ignore the dashed line, which is the Bayes decision boundary.)



- ☐ $k = 1$
- ☐ $k = 2$
- ☒ $k = 10$
- ☐ $k = 100$

[3 pts] Consider training a decision tree given a design matrix $X = \begin{bmatrix} 6 & 3 \\ 2 & 7 \\ 9 & 6 \\ 4 & 2 \end{bmatrix}$ and labels $y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$. Let f_1 denote feature 1, corresponding to the first column of X , and let f_2 denote feature 2, corresponding to the second column. Which of the following splits at the root node gives the highest information gain? (Select one.)

- ☐ $f_1 > 2$
- ☐ $f_2 > 3$
- ☒ $f_1 > 4$
- ☐ $f_2 > 6$

[3 pts] In terms of the bias-variance decomposition, a 1-nearest neighbor classifier has _____ than a 3-nearest neighbor classifier.

- ☒ higher variance
- ☐ higher bias
- ☐ lower variance
- ☒ lower bias

[3 pts] The firing rate of a neuron

☐ determines how strongly the dendrites of the neuron stimulate axons of neighboring neurons

☐ only changes very slowly, taking a period of several seconds to make large adjustments

☒ is more analogous to the output of a unit in a neural net than the output voltage of the neuron

☐ can sometimes exceed 30,000 action potentials per second