10-601 Recitation

Mladen Kolar

Topics covered after the midterm

Learning Theory

Hidden Markov Models

Neural Networks

Dimensionality reduction

Nonparametric methods

Support vector machines

Boosting

Learning theory

$$\Pr[(\exists h \in H) s.t. (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)] \ \leq \ |H| e^{-\epsilon m}$$

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Agnostic learning

$$m \geq \frac{1}{2\epsilon^2}(\ln|H| + \ln(1/\delta))$$

Infinite number of hypothesis

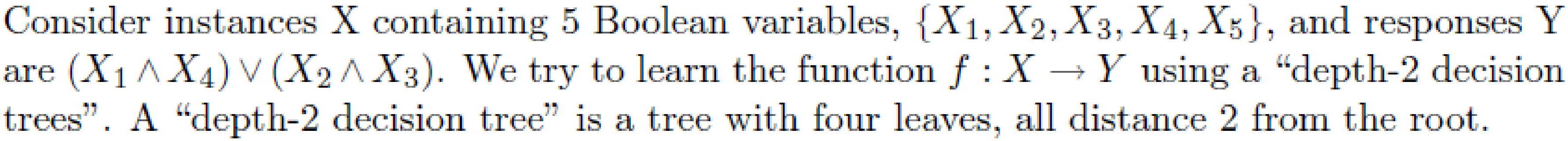
$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

Which formula to use?

Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (X_2 \wedge X_3)$. We try to learn the function $f : X \rightarrow Y$ using a 2-layered neural network.

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

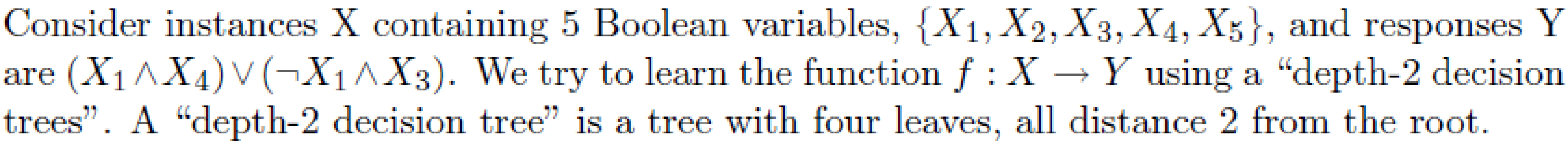
Which formula to use?



Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (X_2 \wedge X_3)$. We try to learn the function $f : X \to Y$ using a "depth-2 decision trees". A "depth-2 decision tree" is a tree with four leaves, all distance 2 from the root.

$$m \geq \frac{1}{2\epsilon^2}(\ln|H| + \ln(1/\delta))$$

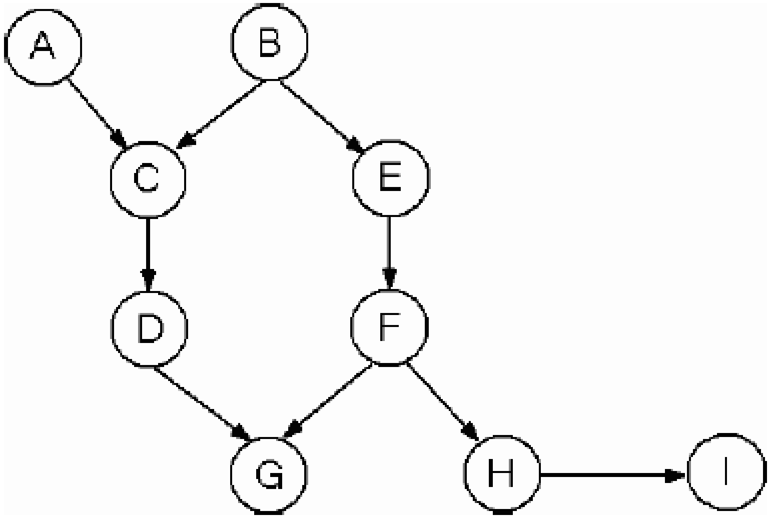
Which formula to use?



Consider instances X containing 5 Boolean variables, $\{X_1, X_2, X_3, X_4, X_5\}$, and responses Y are $(X_1 \wedge X_4) \vee (\neg X_1 \wedge X_3)$. We try to learn the function $f : X \rightarrow Y$ using a "depth-2 decision trees". A "depth-2 decision tree" is a tree with four leaves, all distance 2 from the root.

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

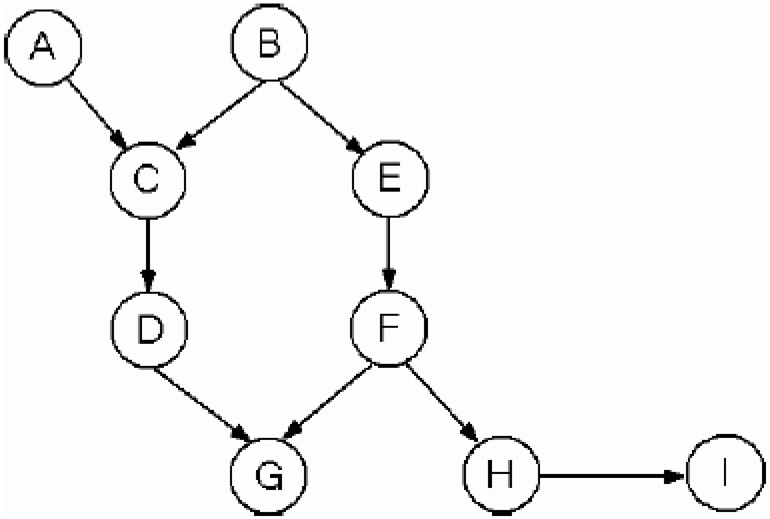
Bayes Networks



Which statement is true?

(a) $P(A, B|G) = P(A|G)P(B|G)$;
(b) $P(A, I) = P(A)P(I)$;
(c) $P(B, H|E, G) = P(B|E, G)P(H|E, G)$;
(d) $P(C|B, F) = P(C|F)$.

Bayes Networks



How many independent parameters?

Hidden Markov Models

Special case of Bayes Networks

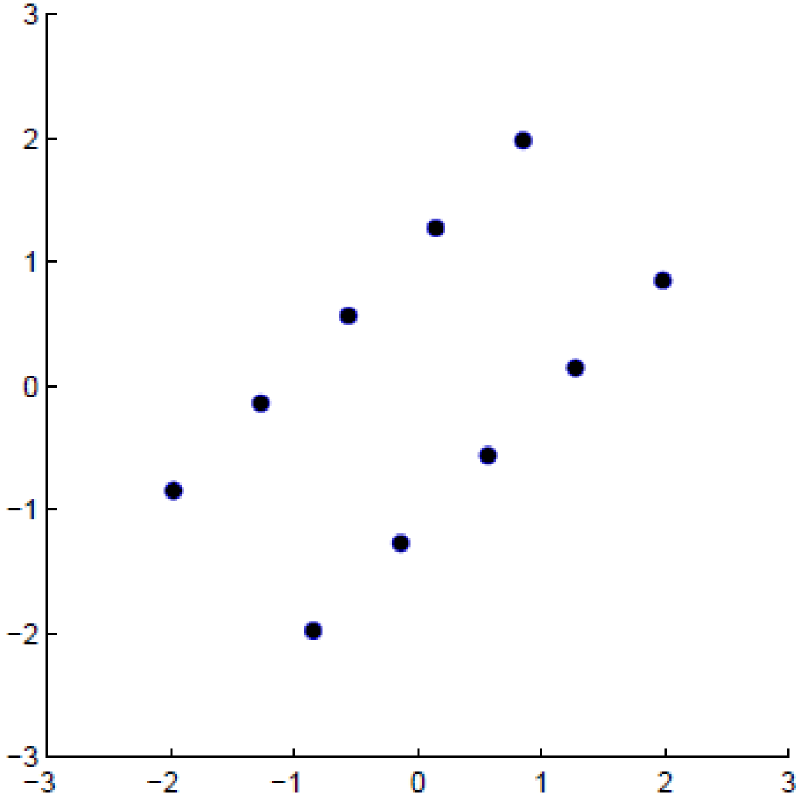
Representation

Algorithms

- forward
- forward-backward
- viterbi
- Baum-Welch

Principal component analysis

Classification in lower dimensional space.

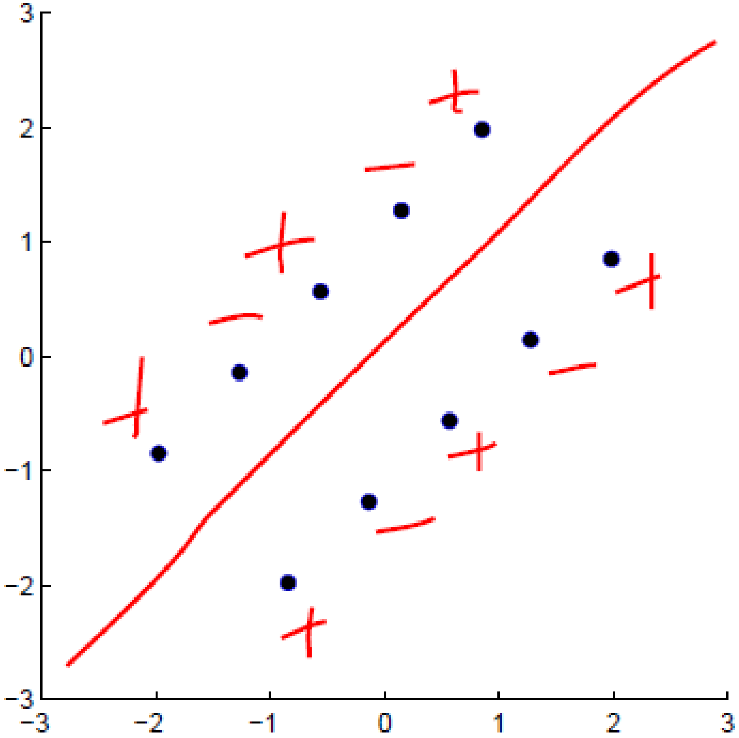


Label points so that 1-NN classifier have the following leave-one-out cross validation errors

2D data: 100% error
1D data from PCA: 0% error

Principal component analysis

Classification in lower dimensional space.

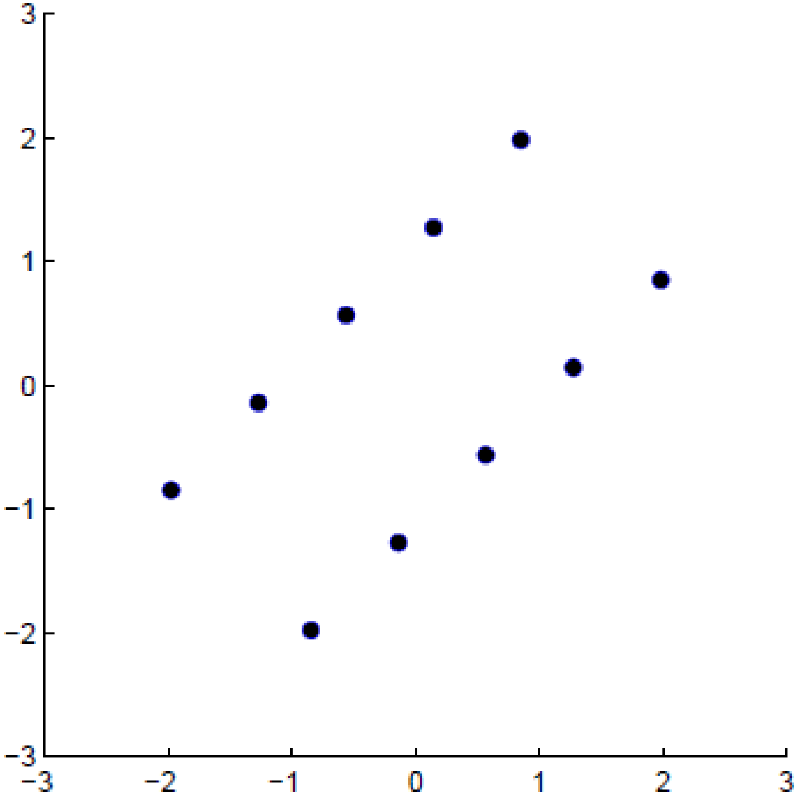


Label points so that 1-NN classifier have the following leave-one-out cross validation errors

2D data:	100% error
1D data from PCA:	0% error

Principal component analysis

Classification in lower dimensional space.

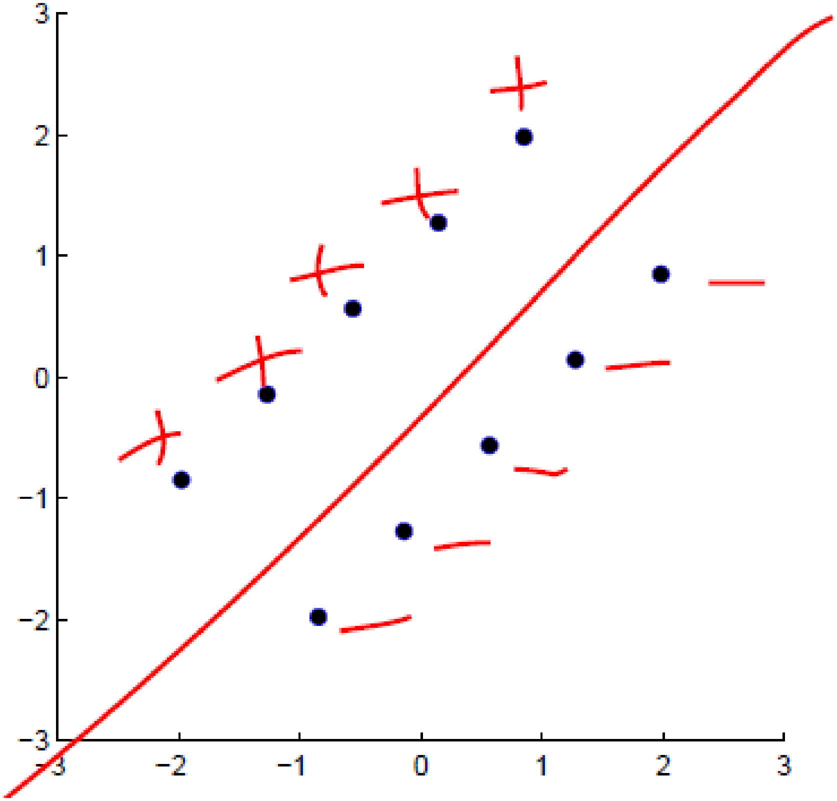


Label points so that 1-NN classifier have the following leave-one-out cross validation errors

2D data	0% error
1D data from PCA:	100% error

Principal component analysis

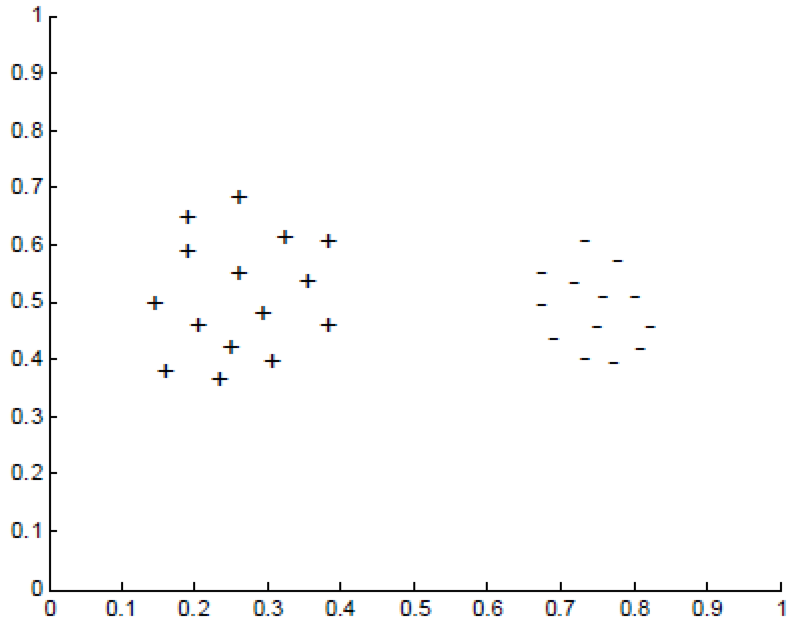
Classification in lower dimensional space.



Label points so that 1-NN classifier have the following leave-one-out cross validation errors

2D data	0% error
1D data from PCA:	100% error

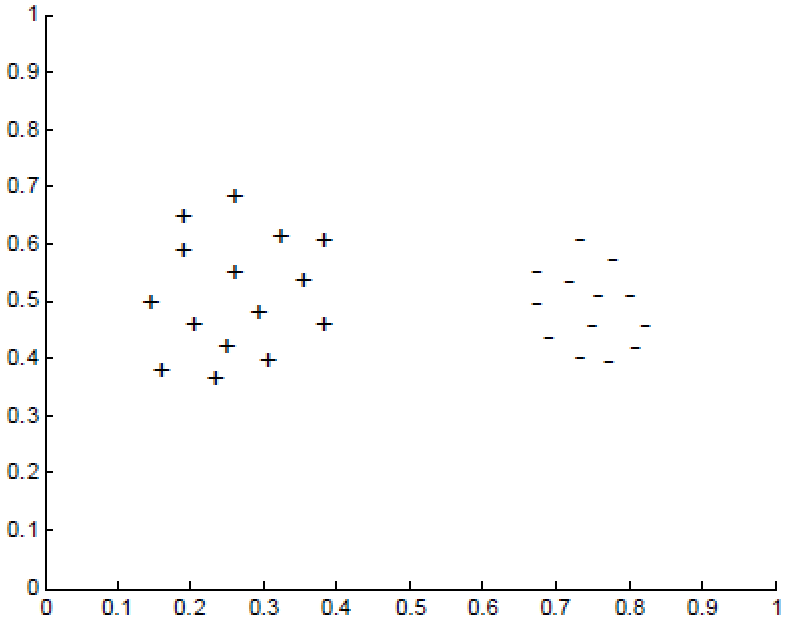
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: T F

LR can separate: T F

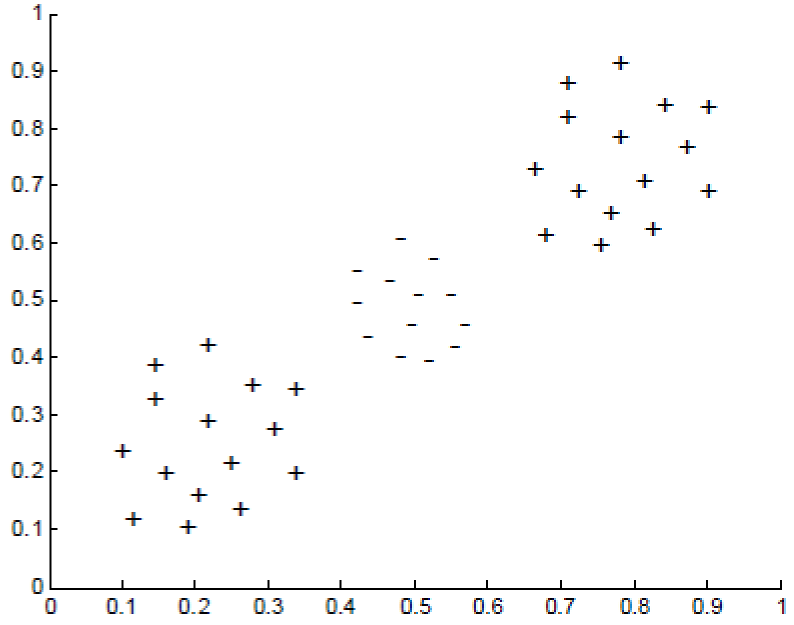
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: [✓ T] F

LR can separate: [✓ T] F

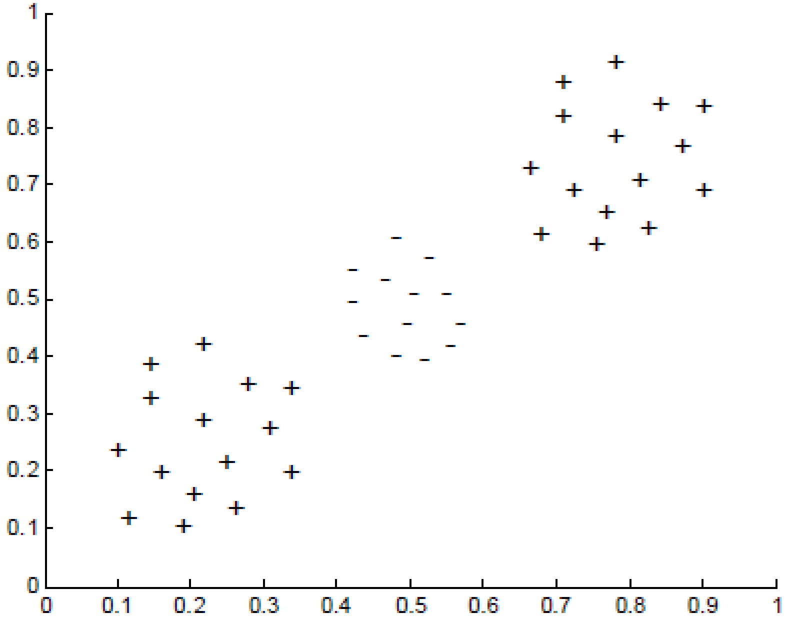
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: T F

LR can separate: T F

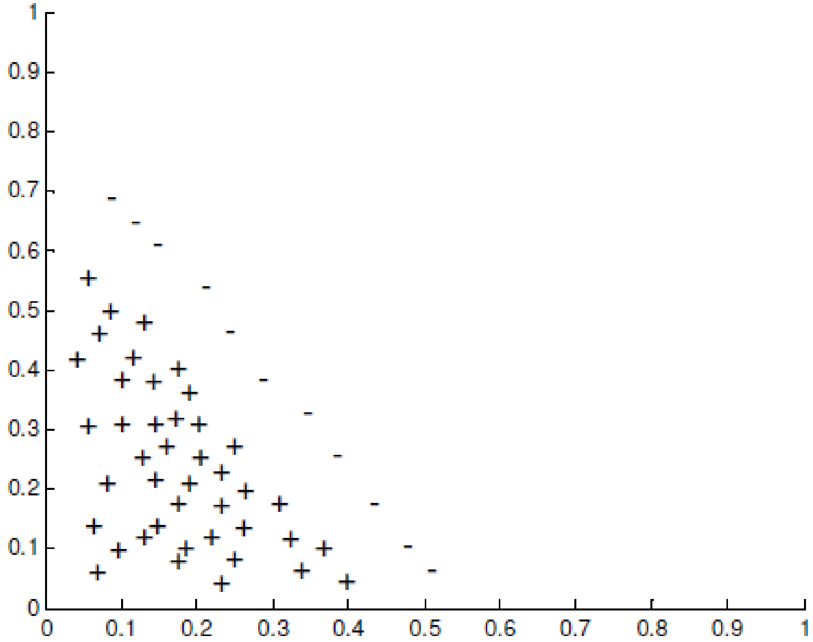
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: [✓ T] F

LR can separate: T [✓ F]

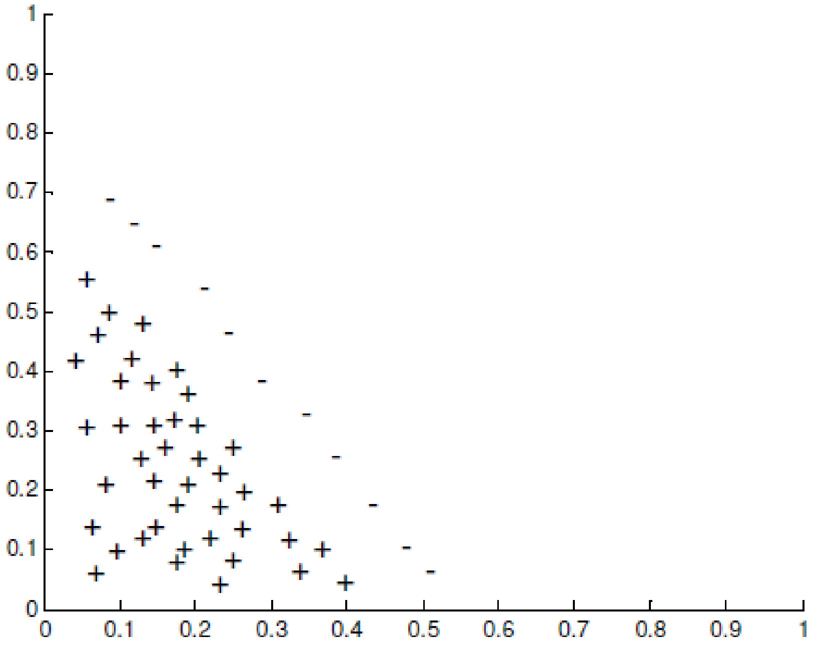
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: T F

LR can separate: T F

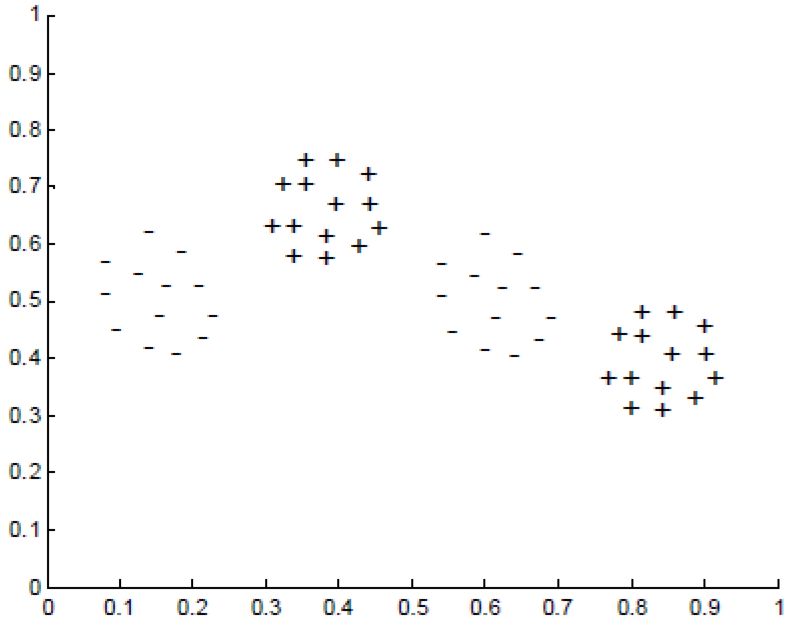
Gaussian Naïve Bayes and Logistic Regression



GNB can separate: T [✓ F]

LR can separate: [✓ T] F

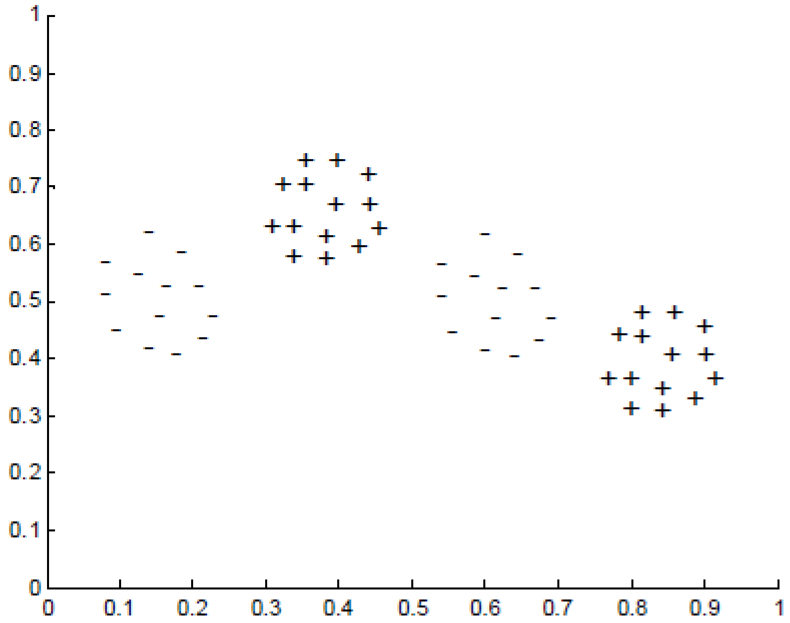
Gaussian Naïve Bayes and Logistic Regression



GNB can separate : T F

LR can separate: T F

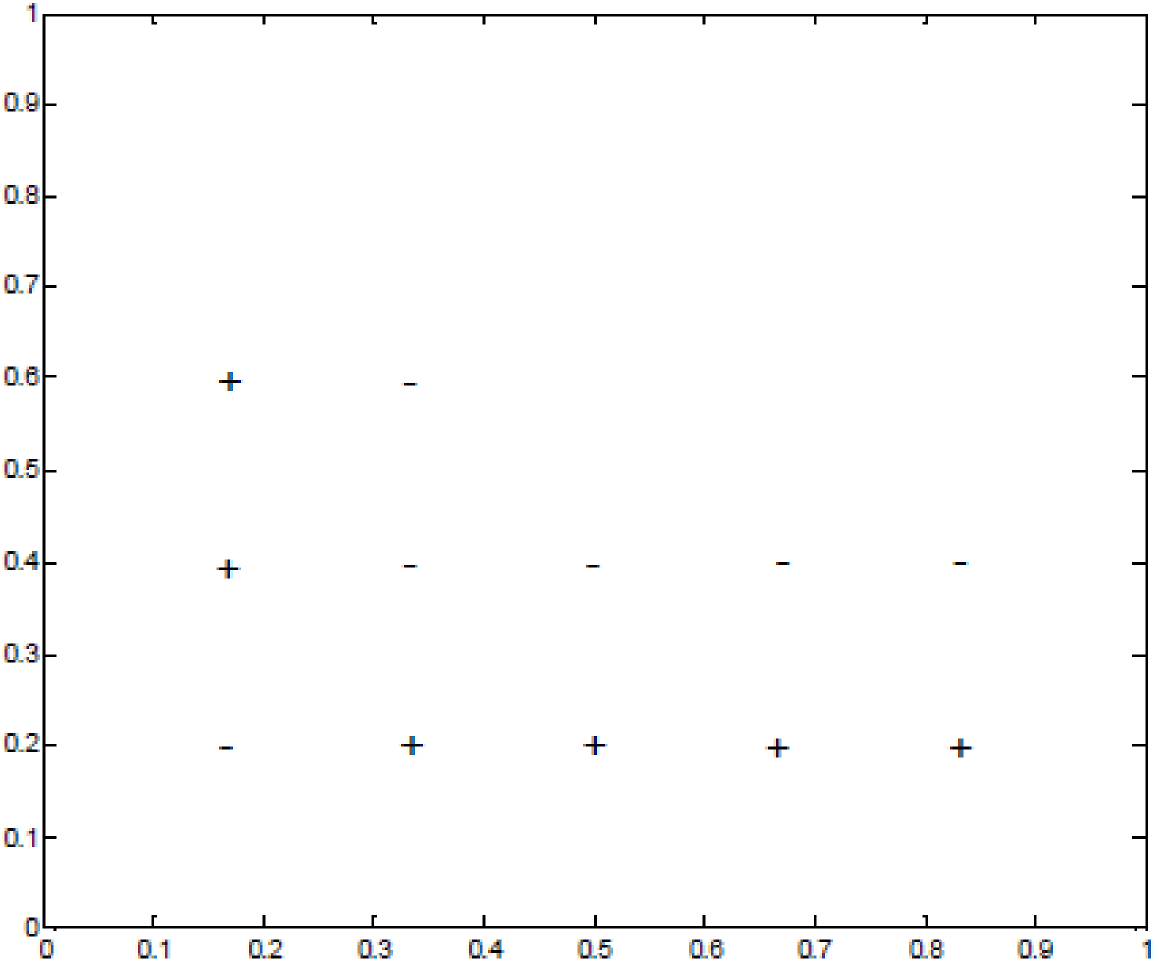
Gaussian Naïve Bayes and Logistic Regression



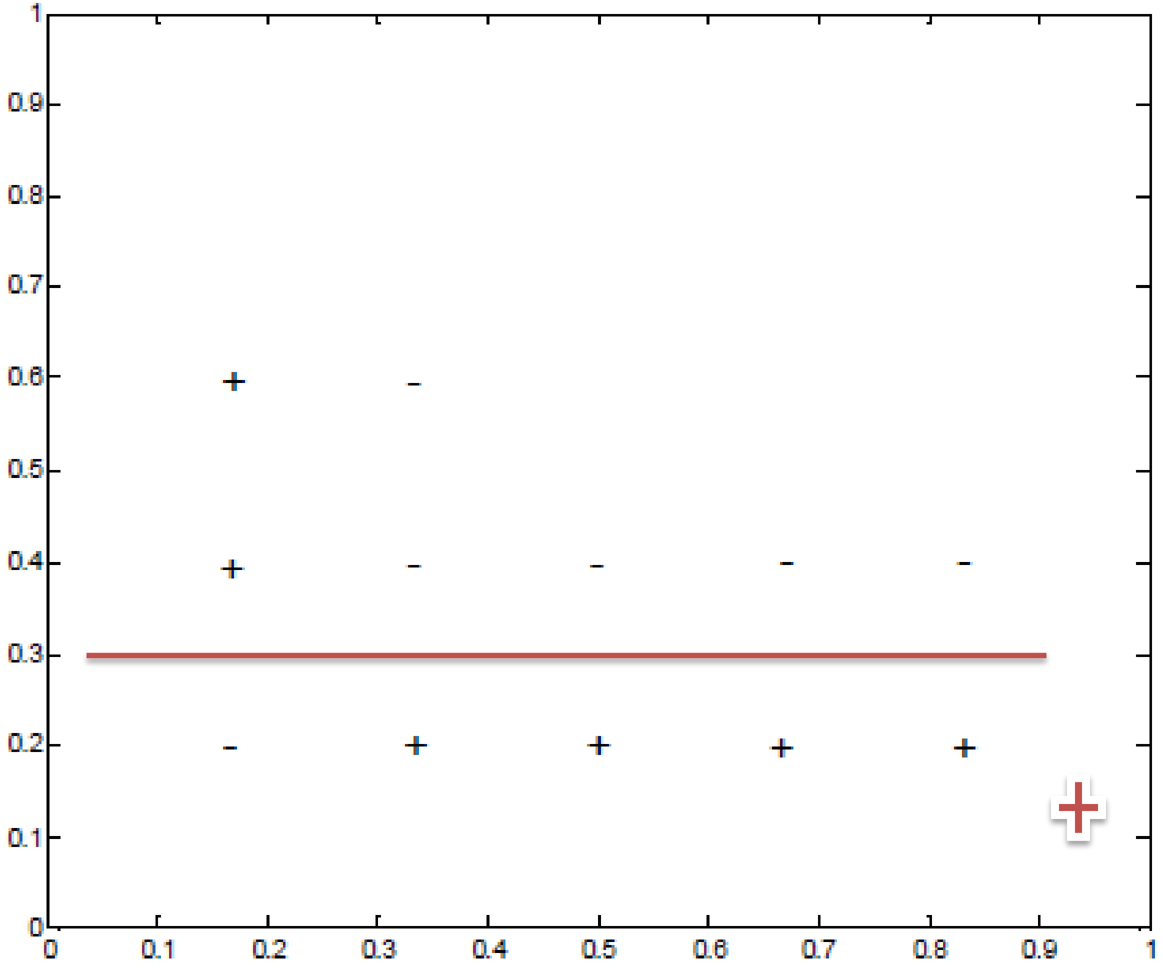
GNB can separate : T [✓ F]

LR can separate: T [✓ F]

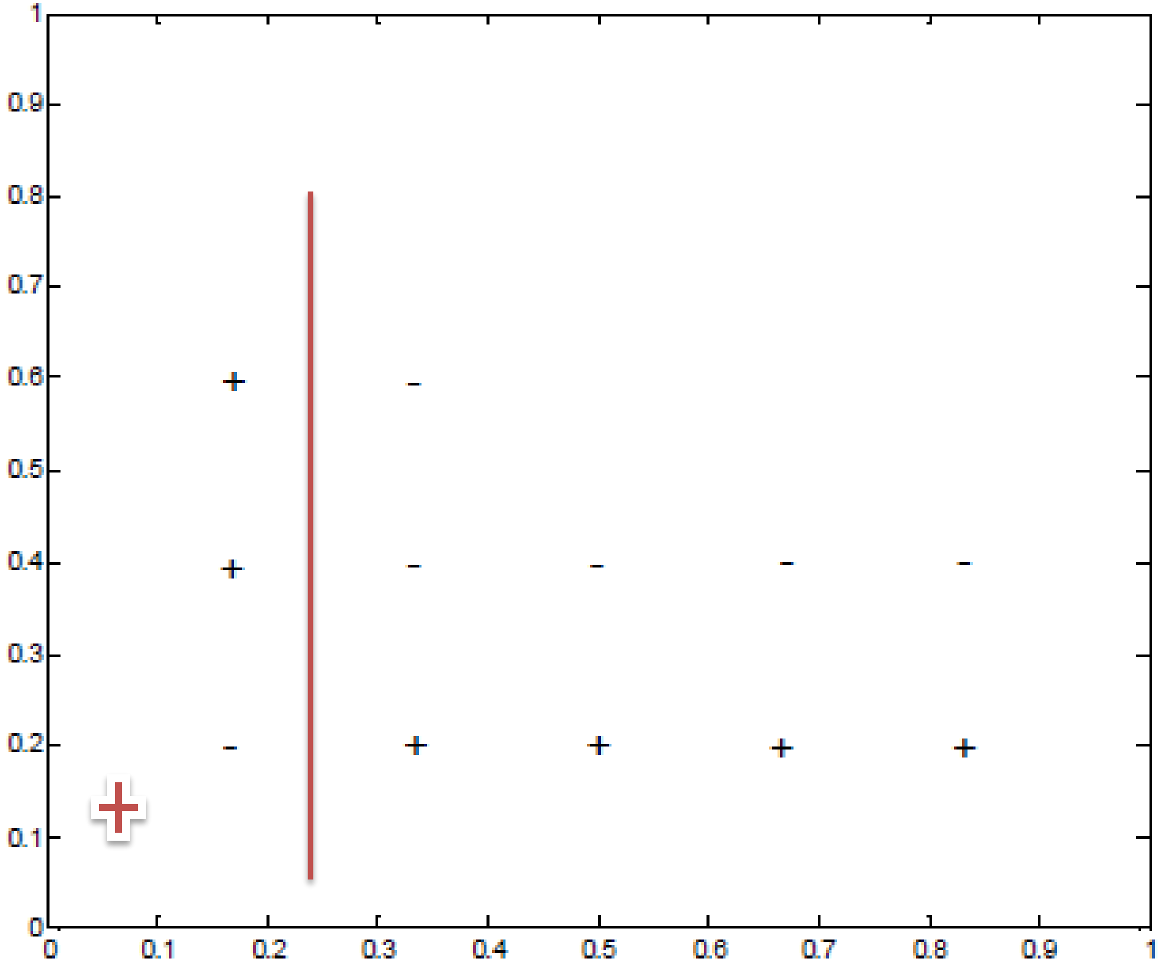
Boosting



Boosting



Boosting



Which classifier to use?

Your billionaire friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative or generative classifier? Why?

Which classifier to use?

Your billionaire friend also wants to classify companies to decide which one to acquire. This project has lots of training data based on several decades of research. To create the most accurate classifier, do you recommend using a discriminative or generative classifier? Why?

Short questions

Assume that we are using an SVM classifier with a Gaussian kernel. Draw a graph showing two curves: training error vs. kernel bandwidth and test error vs. kernel bandwidth

Short questions

Assume that we are modeling a number of random variables using a Bayesian Network with n edges. Draw a graph showing two curves: Bias of the estimate of the joint probability vs. n and variance of the estimate of the joint probability vs. n.

Short questions

Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors. Explain which error is being minimized in each algorithm.

Short questions

Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors. Explain which error is being minimized in each algorithm.

PCA – reconstruction error

Linear regression – residual error

Short questions

For kernel regression, which one of these structural assumptions is the one that most affects the trade-off between underfitting and overfitting:

Kernel function (say, Gaussian vs box-shaped)

Metric used (L_2 vs L_1 vs L_∞)

[✓ Kernel width]

The maximum height of a kernel

Short questions

Given two classifiers A and B, if A has a lower VC-dimension than B then A almost certainly will perform better on a test set.

False

Short questions

In neural networks, what controls the trade-off between underfitting and overfitting?

Short questions

Distribute the numbers 1,2,3,4 to the following four methods for classifying 2D data, such that 1 indicates highest variance and lowest bias, and 4 indicates lowest variance and highest bias.

- Boosted logistic regression (10 rounds of boosting)
- 1-nearest neighbor classifier
- Decision trees of depth 10
- Logistic regression

Short questions

Distribute the numbers 1,2,3,4 to the following four methods for classifying 2D data, such that 1 indicates highest variance and lowest bias, and 4 indicates lowest variance and highest bias.

1. 1-nearest neighbor classifier
2. Decision trees of depth 10
3. Boosted logistic regression (10 rounds of boosting)
4. Logistic regression

Short questions

The ID3 algorithm is guaranteed to find the optimal decision tree.

False

Short questions

What is the difference between MLE and MAP estimates?

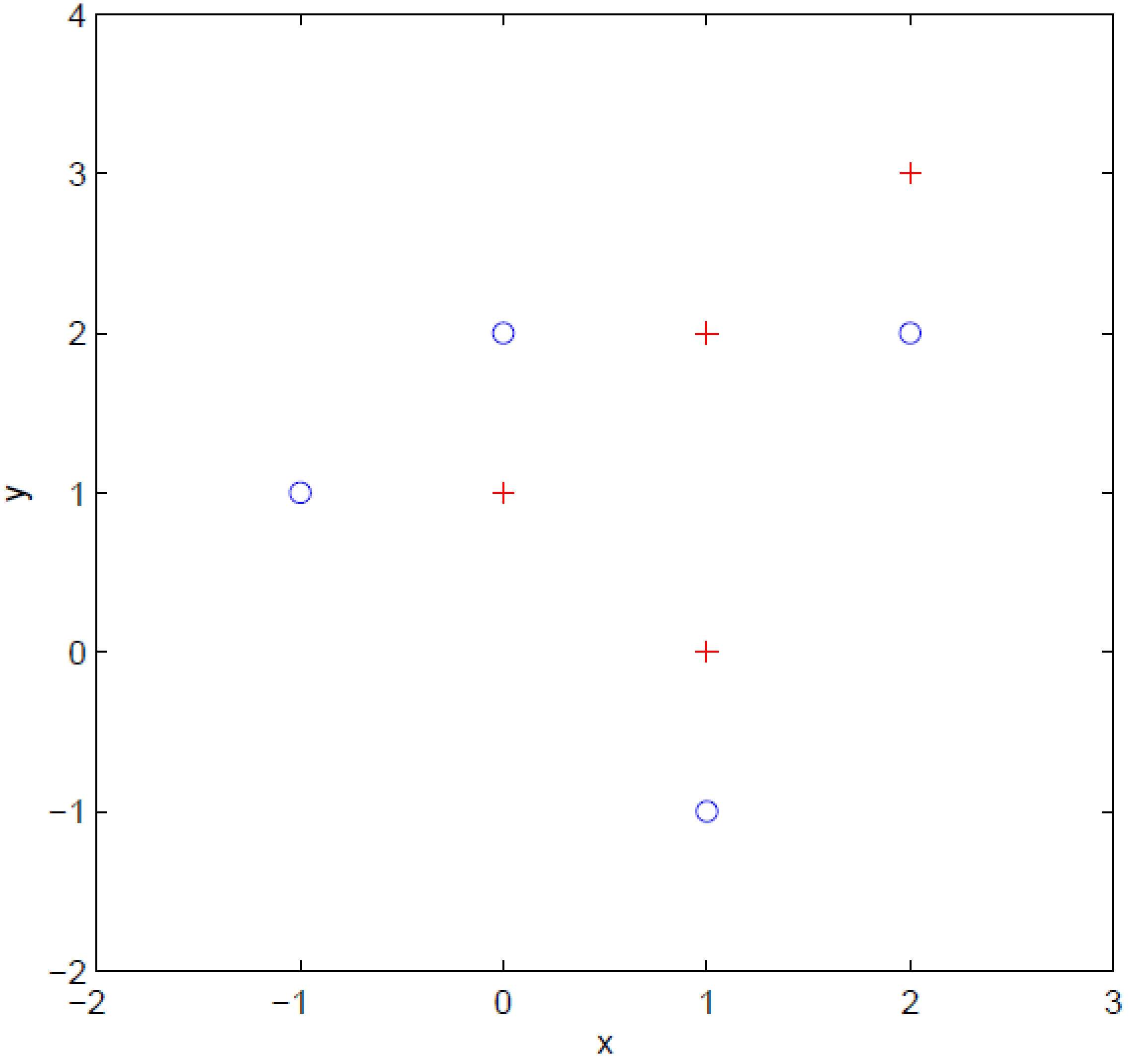
MLE finds parameters to maximize the likelihood function.

MAP finds parameters to maximize the posterior probability.

Is HMM appropriate model?

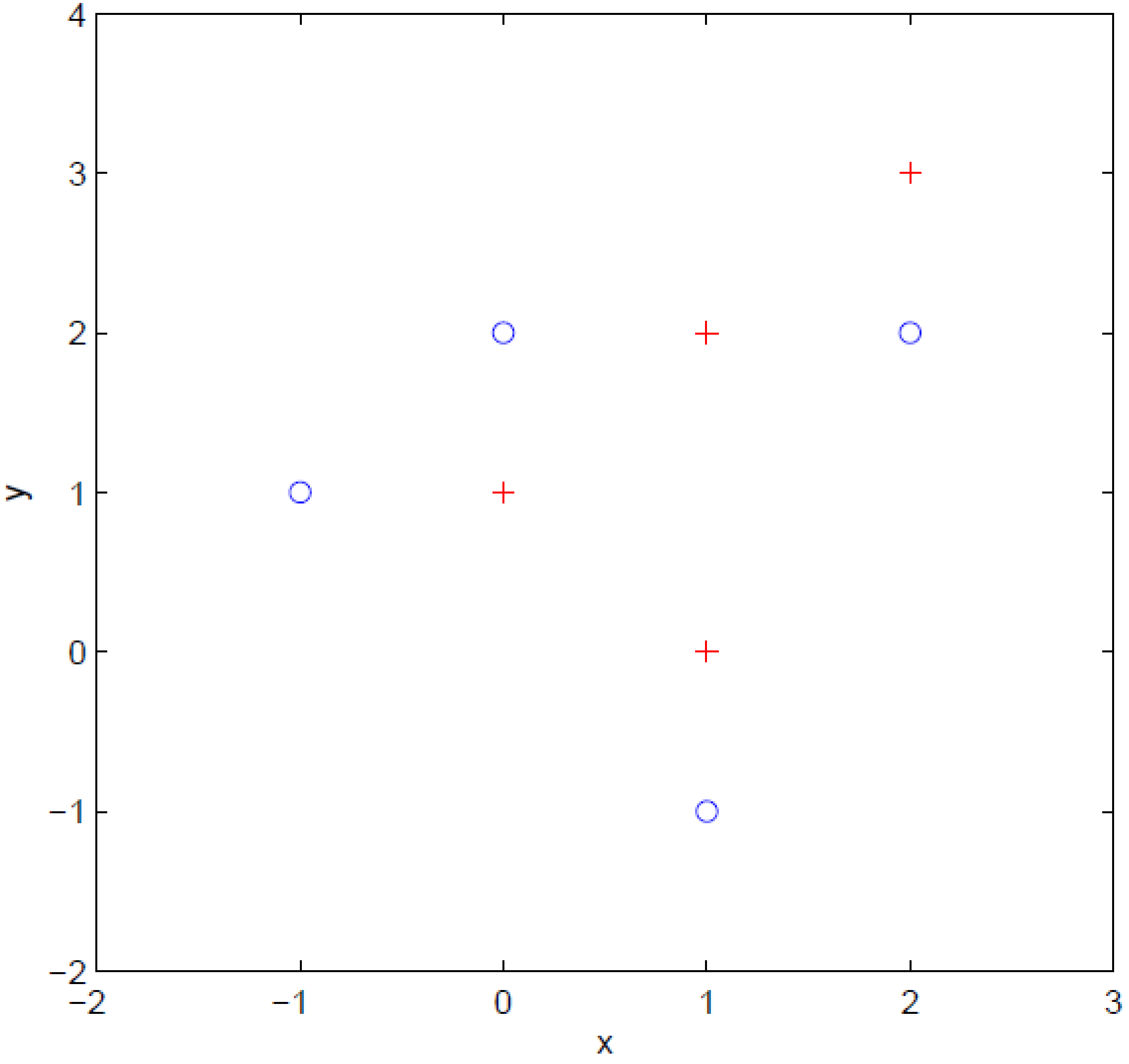
- Gene sequence dataset ←
- A database of movie reviews
- Stock market price dataset ←
- Daily precipitation data from the Northwest of the US ←

kNN



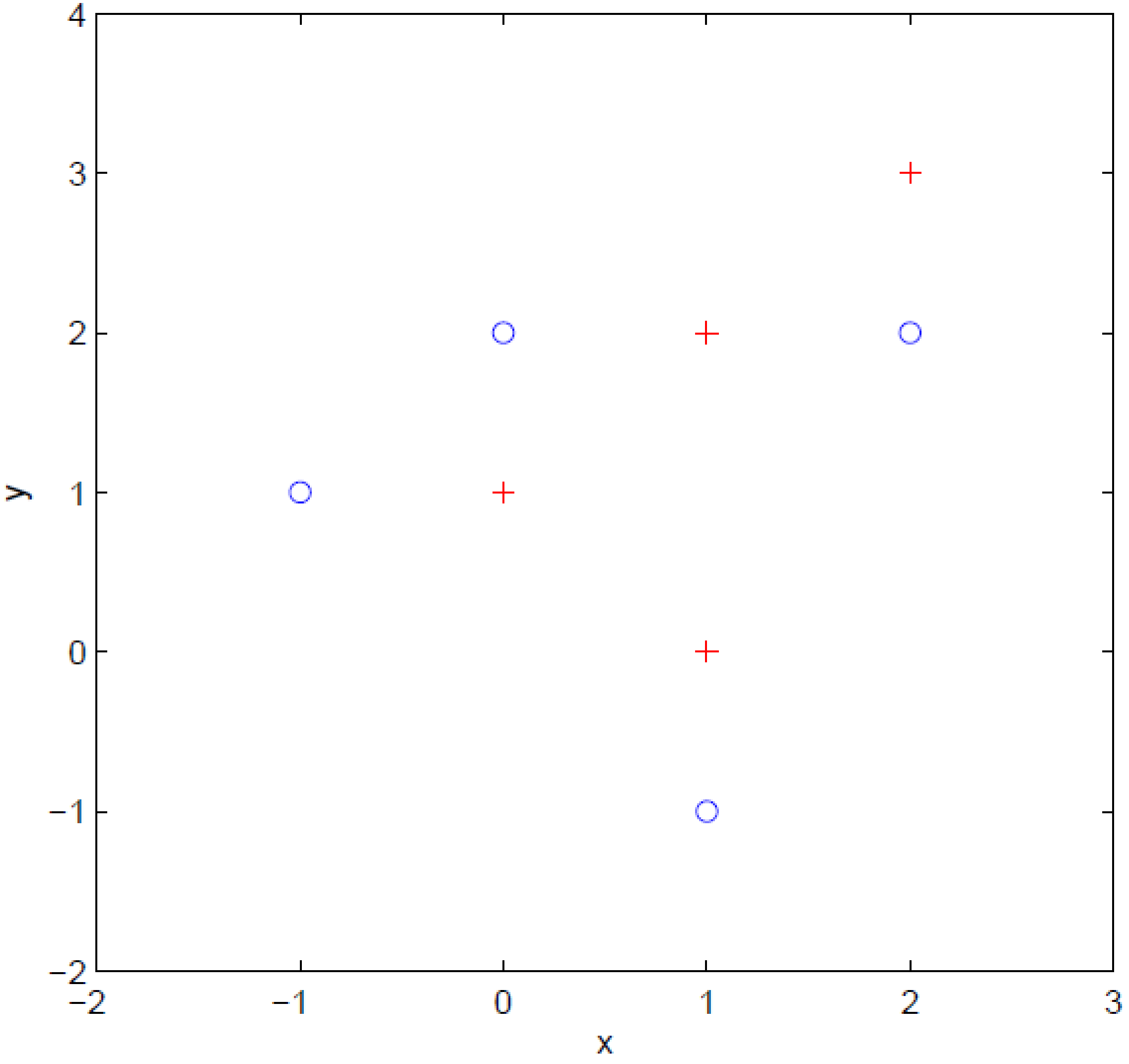
What is the prediction of the 3-nearest-neighbor classifier at the point (1,1)?

kNN



What is the prediction of the 5-nearest-neighbor classifier at the point (1,1)?

kNN



What is the prediction of the 7-nearest-neighbor classifier at the point (1,1)?

kNN

Consider the two-class classification problem. At a data point *x*, the true conditional probability of a class *k is* $p_k(x) = P(C = k|X = x)$.

What is the Bayes error at a point x?

$$\min_k p_k(x)$$

What is the 1-NN error at x when x’ is the nearest neighbor?

$$p_0(x)p_1(x') + p_0(x')p_1(x)$$