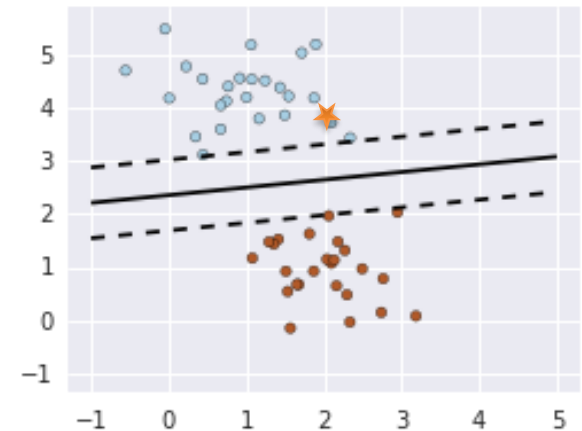
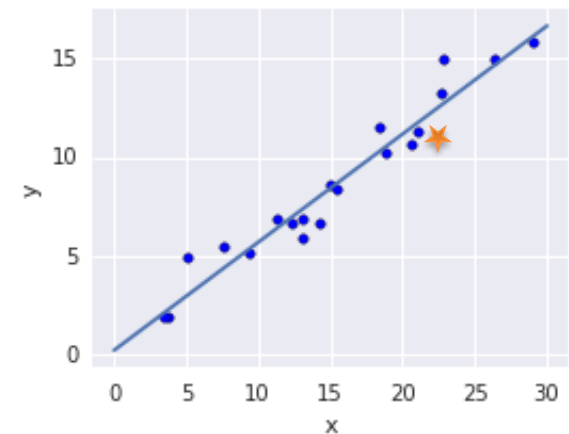


- Machine learning explores the study and construction of algorithms that can learn from and make predictions on data
- Unsupervised learning
 - clustering
 - dimensionality reduction
- Supervised learning
 - regression (label is continuous)
 - **classification** (label is discrete)

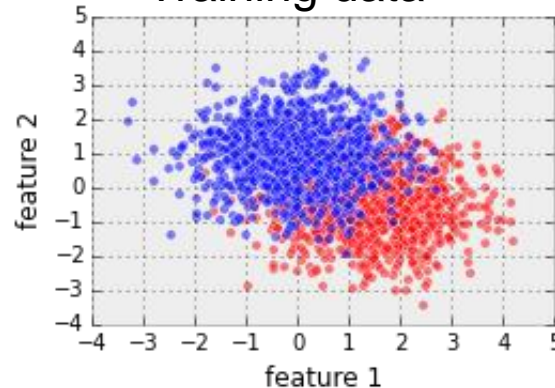


Machine learning – training and validating a binary classification model

Point:

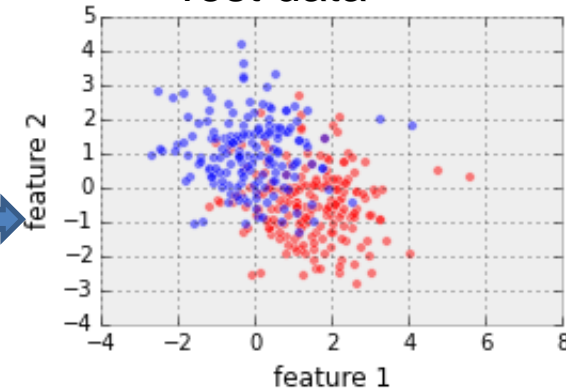
- Sample with features 1 and 2
- Color indicates class

Training data



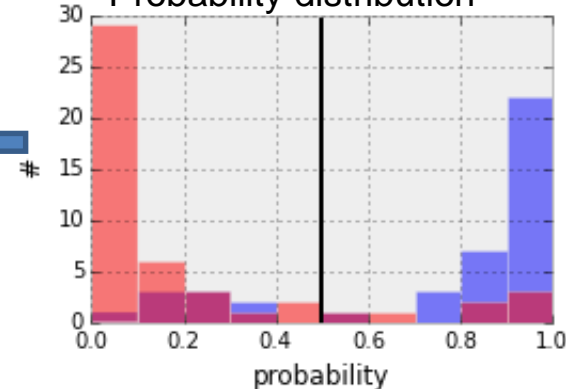
Train your model

Test data

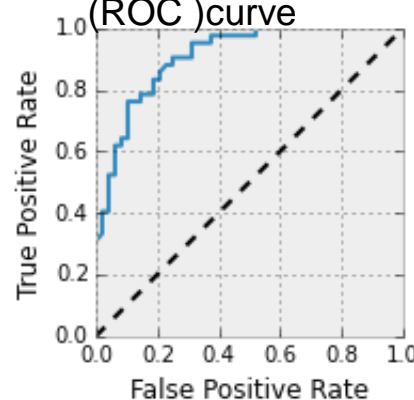


Make predictions on test set

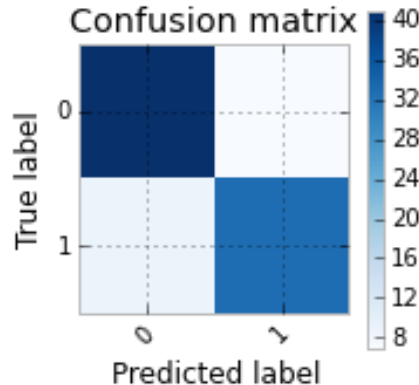
Probability distribution



Receiver Operating Characteristic (ROC) curve

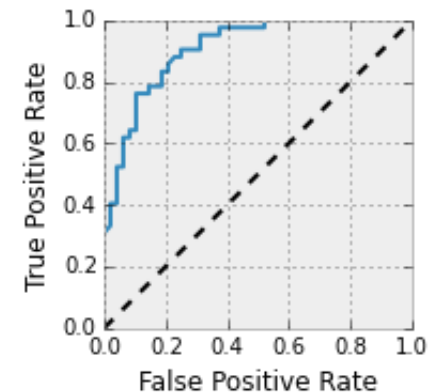


Confusion matrix



- Accuracy: fraction of correct predictions
- Are under the ROC curve (AUC)
 - 0.5 – performance of random selection
 - 1 – ideal classifier
 - > 0.9 diagnostic tests in medicine
- Confusion matrix
 - Sensitivity
 - Recall of positive, TP fraction
 - $TP / (TP + FN)$
 - Specificity
 - Recall of negative, TN fraction
 - $TN / (TN + FP)$

Receiver Operating Characteristic (ROC) curve



		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

TP – True positive, TN – True negative
FP – False positive, FN – False negative

Machine learning in Python

Scikit-learn

scikit-learn algorithm cheat-sheet

