TP 1 - Apprentissage Statistique Appliqué

Nokri Amale, Rahis Erwan, Vuillemot Bertrand

2020 - 2021

Contents

1 Partie 1

1.1 Cross-Validation with GridSearchCV

Question : Explain in your report what happens when we run clf.fit(X_train, Y_train). What is the complexity for each of the three following cases?

Answer : The line clf.fit(X_train, Y_train) here uses the fit function on the object named clf which is an object of the class GridSearchCV. This function will fit the parameters of the clf object which is taking as parameter an object named knn of the class KNeighborsClassifier(), a dictionary named parameters containing the number of neighbors to be tested in the knn algorithm (5 here) and the cv parameter referring to the number of folds to be used in the cross-validation. Basically it will perform a 3-folds cross-validation on a kNN model with 1 to 5 neighbors on the train sample and it will allow us to have the best model. They are all part of the sklearn package.

Question : What is the test accuracy? What would be the accuracy of random guess?

Answer : The test accuracy is the measure of how often the points are correclty classified. In our case the accuracy is 0.875. It means that 87.5% of the time, the points are correctly classified on the test sample. If we did a random guess we would randomly choose an output in the range 0 to 9 so the accuracy would converge towards $\frac{1}{10}$ according to the LLN.

Question : What is LinearSVC() classifier? Which kernel are we using? What is C? (this is a tricky question, try to find the answer online)

Answer : LinearSVC means Linear Support Vector Classification. We are using a linear kernel. The parameter C represents the regularization weights, ie the penalty applied on the loss function. The loss function used here is the Squared Hinge Loss : $l(y) = \max(0, 1 - t \cdot y)$

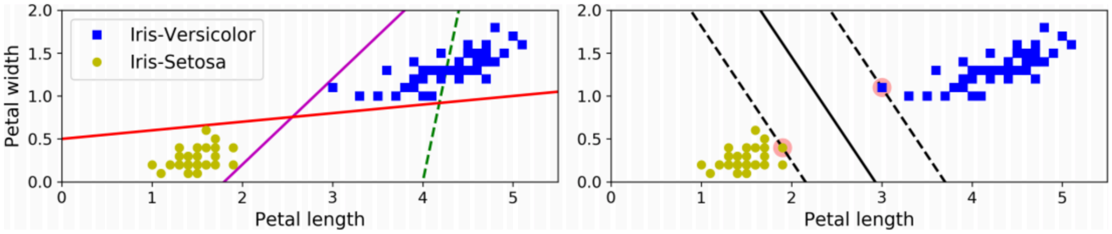

Figure 1: Example SVM

Question : What is the outcome of np.logspace(-8, 8, 17, base=2)? More generally, what is the outcome of np.logspace(-a, b, k, base=m)?

Answer : The logspace function from the numpy package will return k numbers going from -a to b on a log scale with a log base m.
Answer : The outcome of np.logspace(-8, 8, 17, base=2) is a logarthmic space going from 2^{-8} to 2^{8} with 17 numbers equally spaced on log scale. The logspace function from the numpy package will return k numbers going from m^{-a} to m^{b} spaced on a log scale with a log base m.

Question : What is the meaning of the warnings? What is the parameter responsible for its appearence?