CHALLENGE 2

Riccardo Striano

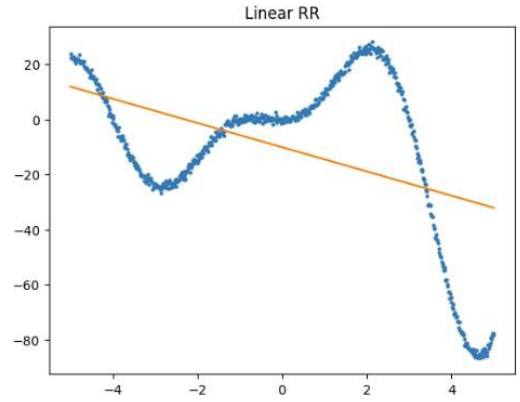
INTRO

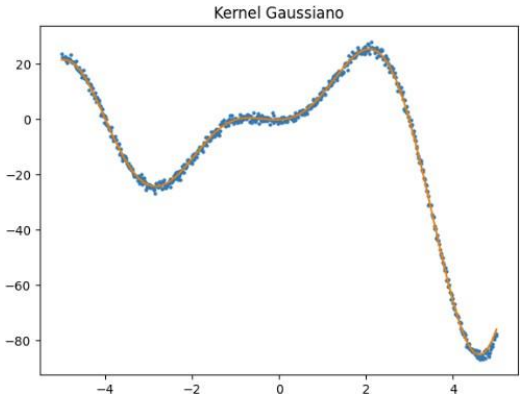
Lo scopo della challenge era investigare i vantaggi che i kernel possono dare per la Ridge Regression e la PCA per dati non lineari o classi non linearmente separabili.

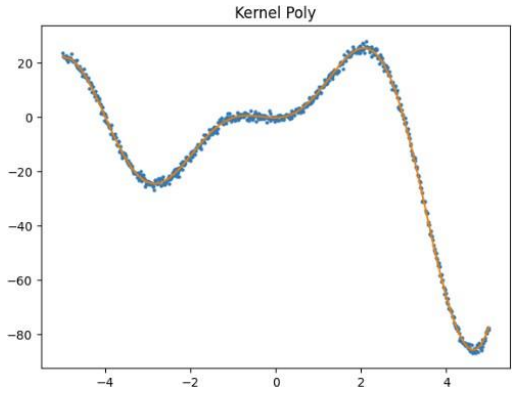
RIDGE REGRESSION

La Ridge Regression lineare fitta molto poveramente i dati. Sia il Kernel Gaussiano che Polinomiale invece funzionano molto bene, con quest'ultimo come migliore.

RMSE = 26.80 RMSE = 0.34 RMSE = 0.16

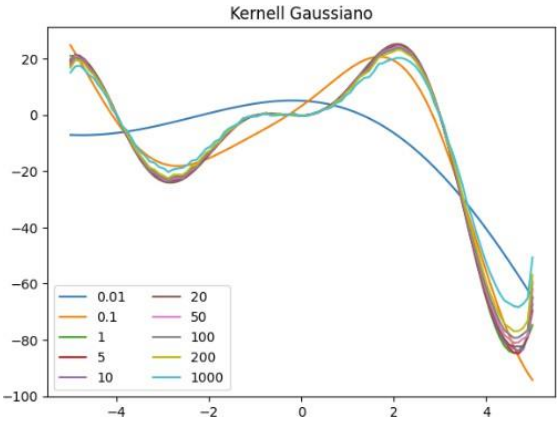




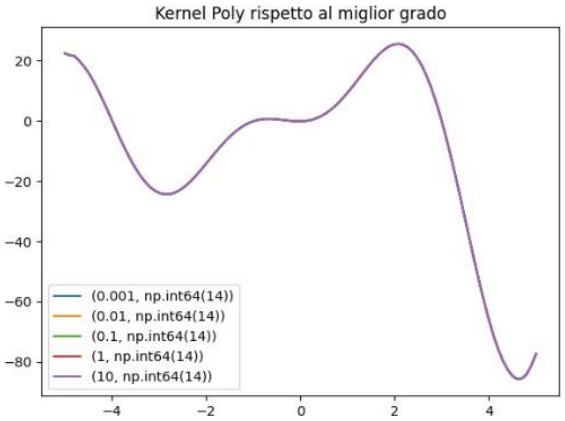


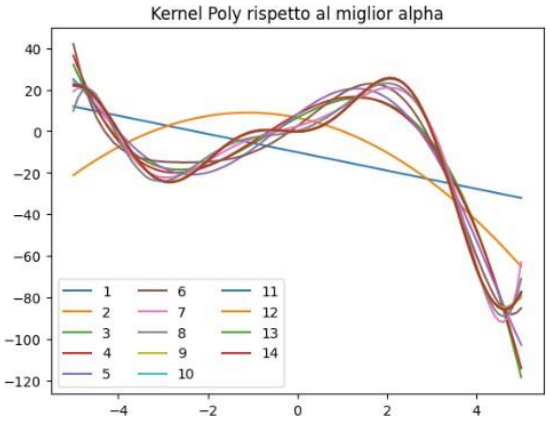
-SCELTA DEI PARAMETRI

- Per il Kernel Gaussiano il parametro migliore è risultato gamma = 1



- Per il Kernel Polinomiale invece abbiamo alpha = 0.001 e grado = 14



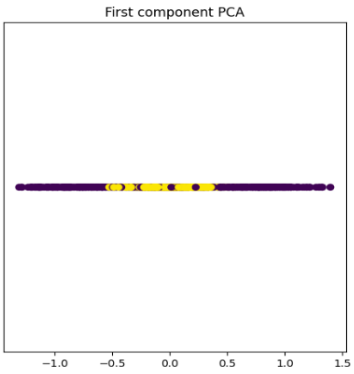


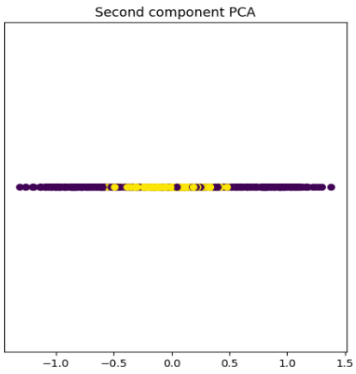
Notiamo come una volta individuato un grado abbastanza buono alpha non ha praticamente più effetto.

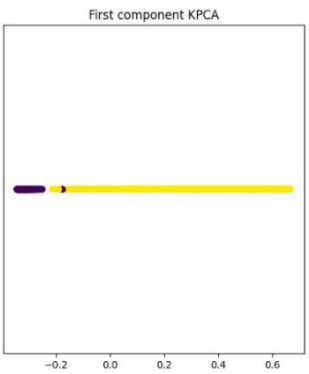
PCA

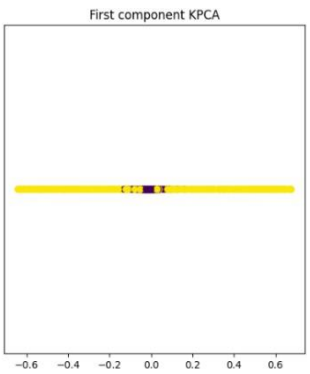
Anche qui la Kernelizzazione del metodo porta a risultai molto migliori.

-Circles dataset:
Il circles datatset presenta 2 cerchi concentrici ovviamente non separabili linearmente.
PCA infatti non riesce a trovare delle componenti principali che siano separabili linearmente con una SVM.
Kernel PCA invece ci riesce molto bene. (come vediamo sia dai grafici sotto che dallo score degli SVM)
Gli SVM performano con una accuracy del 49,6% per PCA, contro un 99,2% per Kernel PCA.









-General non linear dataset:

Ho quindi cercato di confermare la generalità della conclusione trovata, cioè che i kernel ci permettono di fare regressione e separazione di dati non lineari, performando una serie di PCA vs Kernel PCA test su una serie di dataset generati casualmente grazie al comando *“sklearn.datasets.make_classification()”.* Andando poi a vedere le accuracy medie di un modello SVM trainato con le feature estratte dall’una e dall’altra.

Quello che notiamo è che la PCA normale ha una accuracy media dell’71.21% mentre la Kernel PCA del 85.2%. Una seconda prova ha portato ad una accuracy media dell’82% per PCA e 91% per Kernel PCA .La differenza è chiara e notevole.

CONCLUSIONI

I modelli con kernel mostrano chiaramente un miglioramento rispetto al modello lineare sia per quanto riguarda la regressione che la classificazioni. Ciò è ovvio quando consideriamo il fatto che i dati che prendiamo in considerazione non seguono trand lineari ne sono linearmente separabili