

ITMAL Øvelser – Regression

Øvelse 1 :

I skal afprøve lineær regression på Diabetes datasæt i Scikit Learn (OBS: Start med at se "tips" nedenunder).

- a) Plot "Disease progression" (target værdi – mål for hvor meget sygdommen har udviklet sig, dvs. "hvor syg man er") i forhold til de første fire features (alder, køn, bmi og blodtryk), hver for sig – dvs. 4 plots. Ser I en sammenhæng mellem features og target ?
- b) Prædiker Disease Progression ud fra kun 1 feature (den som ser ud til at passe bedst i sprm a) – gøres ved brug af "fit" med linear regression model. Hvor stor er fejlen ? (root-mean-square fejl).
- c) Lav et plot med både data og prædiktions-linien – ser modellen ud til at fitte godt til data ?
- d) Plot et histogram over residualerne – er de normalfordelte ? (Dette fortæller noget om data – er der fx. outliers, som kan være meget generende for vores model, eller som er decideret fejlmålinger)
- e) Prædiker Disease Progression ud fra de første 4 features. Find root mean squared error (RMSE) og sammenhold med sprm. b
- f) Prøv også med alle features. Er RMSE bedre ?
- g) Afprøv samme metode, men nu med et train/test split (fx. 70/30) – Har I overfittet i sprm. f ? (relater til antal parametre i jeres model og antal samples i datasættet)

Tips :

Datasæt beskrivelse - https://scikit-learn.org/dev/datasets/toy_dataset.html#diabetes-dataset

Load datasæt - https://scikit-learn.org/dev/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes

Eksempel med lineær regression - https://scikit-learn.org/dev/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py

Øvelse 2 :

I skal afprøve logistisk regression på Height-Weight datasæt (højde og vægt på hhv. mænd og kvinder) – se "read_height_weight_dataset.py".

- a) Først skal I afprøve på kun højde data. I skal plotte højde data for mænd (kategori 0) og kvinder (kategori 1) og dernæst fitte logistisk regression til data. I skal plotte den logistiske funktion.
- b) Dernæst skal I fitte til 2D data og plotte "højde-kurven" for den logistiske funktion (OBS: Når den er lig 0.5, har I decision boundary – dvs. begge klasser lige sandsynlige).
- c) Lav en rigtig træning med train/test sæt split. Find train og test error rates.

Tips :

Fit logistisk funktion - https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression

Contour plot - https://matplotlib.org/gallery/images_contours_and_fields/contour_demo.html#sphx-glr-gallery-images-contours-and-fields-contour-demo-py

Generel surface plot - https://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html