Exercises for
Pattern Analysis Programming
Lina Felsner, Mathias Seuret, Dalia Rodriguez-Salas
Task 4: May 13 – May 20

Pattern Recognition Lab

Transforming an Unsupervised Task to a Supervised Task

Preliminary Remarks: Density Estimation as a Regression Task
There are unsupervised models that operate on the empirical probability density distribution of the data, and there are supervised models to perform classification and regression. Supervised methods can be further split up into a generative probability model for the joint density $p(\mathbf{x}, \mathbf{y})$ of features $\mathbf{x}$ and labels $\mathbf{y}$, or a discriminative model that directly target at $p(\mathbf{y}|\mathbf{x})$.

There is a good reason to have specialized methods for different learning tasks: task-specific performance. However, from a theoretical perspective, it is quite interesting to note how closely related the seemingly disparate tasks of unsupervised learning, regression, and classification are.

This exercise is loosely inspired by Sec. 14.2.4 in "The Elements of Statistical Learning" by Hastie, Tibshirani, Friedman: let us convert the density estimation task into a regression task.

The key idea is to define a second density that models the background. Assign high values to the actual density samples (e.g., the ones from our raccoon), and low values to the background density. Then, the regression task is to fit one function that represents both distributions, with high values in regions where our actual samples dominate, and with low values in regions where the background dominates.

Exercise 1 Reuse the raccoon sampler from the previous exercise to create a density of the samples. We also need a background distribution: what might be a good distribution for representing the background? Hastie, Tibshirani, Friedman suggest that unless we have a better idea, we can certainly choose a uniform distribution, so let us work with a uniform distribution. Note that from a theoretical perspective, you will need at least as many samples in the background distribution as in the foreground distribution.

Use a random forest to fit the regression model. That way, we can also earn some practical experience with parameterization of such a forest. Random forests are implemented in `scikit.learn.ensemble`, specifically in the class `RandomForestRegressor`.

You can check the performance of your regressor visually, by making a regressor prediction for each pixel of the raccoon image: does the output look like a raccoon? A more quantitative approach that also allows to tune the forest parameters would be to reuse the cross-validation from the previous exercise.

Exercise 2 Repeat your experiments using now `ExtraTreesRegressor`. What exactly is the difference between a `RandomForestRegressor` and an `ExtraTreesRegressor`? Play with the amount of samples as well as with the amount of trees and maximum height. Which parameter settings provide satisfying results? By tendency, are the requirements for the amount of training data larger or smaller than for our Parzen window estimator? Choose a set of parameters to create a figure with your results for both regressors (example in Fig. 1). Please post it to the forum

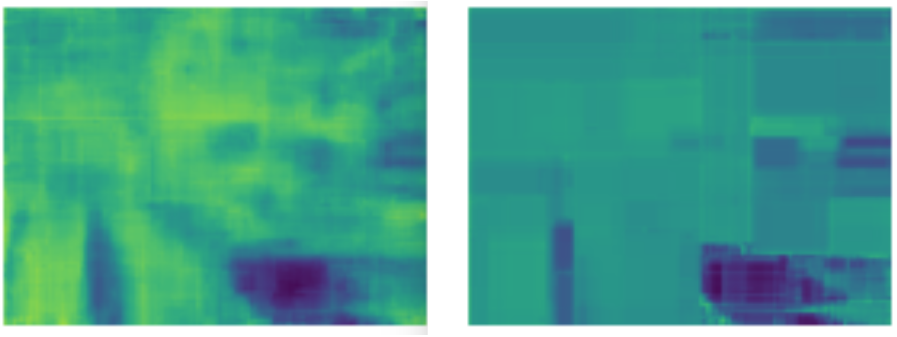

Figure 1: Left side images correspond to ExtraTreesRegressor. Right side images correspond to RandomForestRegressor. Amount of samples = 100000 (50000 foreground + 50000 background), Amount of trees = 20, Maximum depth = 10

and add a short text about one observation you made.

Optional: A Second Dataset
Use the function sklearn.datasets.make moons to create a synthetic den- sity of two interwoven half moons. Repeat your experiments with the RandomForestRegressor and the ExtraTreeRegressor. Are your observations consistent with your observations on the previous dataset? If you like, you can post a figure of the moons instead or additional to the raccoon image to the forum. Please add a short text about one observation you made.

Comments:
We ask for only one figure per group. Please also state your group number. Bring your code to the joint meeting on May 20 (or May 21 (not sure yet, we will keep you informed)) for a potential little extra experiment.