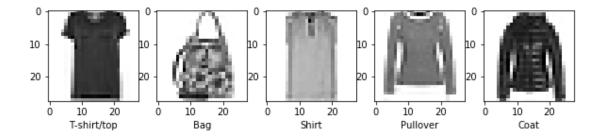# Exercises

Fashion-MNIST

The Fashion-MNIST dataset (https://www.openml.org/d/40996) contains 70,000 images of Zalando fashion products, classified into 10 types of clothing, each represented by 28 by 28 pixel values. You can download it from OpenML and visualize some of the examples

1. Evaluate different models

- Take a *stratified* 10% subsample of the data. Use it for everything below
- Start with k-Nearest Neighbors, Logistic Regression and Linear SVM
- Evaluate these 3 classifiers with their default settings. Use cross-validation with 3 folds, output the accuracy and standard deviation.
- Now compare other models as well (SVM, RandomForest, Boosting,...)

Note: you can use an even smaller sample while testing/bugfixing

2. Tune the parameters using GridSearchCV.

- Vary the main hyperparameter (C or k) for all 3 models. Do the results improve? Which ranges have an effect on performance?
- (Optional) Visualize both the test score and training score as function of the parameters for all three models. Discuss when (for which values) the model is underfitting or overfitting.

Note: you can use a line plot or 1D heatmap for visualization. Use again 3-fold cross-validation and accuracy.

3. Upload your best models to OpenML

- You can use the code below. Enter your OpenML Key.
- Replace the classifier with your best classifier and its hyperparameters
- Run and upload the model
    - Note: since this uses the full MNIST dataset, it will take longer than in exercise 1
- Compare with the other models on https://www.openml.org/t/146825 (https://www.openml.org/t/146825)

4 Analyze which types of misclassifications are made (advanced)

- Create a standard train_test split
- Train Logistic Regression on the training data and generate predictions on the test set
- Visualize (like above) a few examples that are misclassified by Logistic Regression. Discuss: are these indeed difficult cases?
- Build the 'confusion_matrix' of all predictions. Discuss which classes are often confused.

5 Visualize the model parameters (advanced)

- Retrieve the model parameters (coefficients) for LogisticRegression and Linear Support Vector Machines
- Plot the coefficients as a 28*28 image as above.
- Interpret the results. Which pixels does the models pay most attention to? Why? Is there a difference between both models?

6 Try some other datasets

- Browse OpenML for datasets that you find interesting, and try to build models on those.
- Upload your best results to OpenML
- Examples:
  - EEG eye state: data_id:1471 (http://www.openml.org/d/1471), task_id:14951 (http://www.openml.org/t/14951)
  - Volcanoes on Venus: data_id:1527 (http://www.openml.org/d/1527), task_id:10103 (http://www.openml.org/t/10103)
  - Walking activity: data_id:1509 (http://www.openml.org/d/1509), task_id: 9945 (http://www.openml.org/t/9945), 150k instances
  - Covertype (Satellite): data_id:150 (http://www.openml.org/d/150), task_id: 218 (http://www.openml.org/t/218). 500k instances
  - Higgs (Physics): data_id:23512 (http://www.openml.org/d/23512), task_id:52950 (http://www.openml.org/t/52950). 100k instances, missing values