Lab2

Machine Learning 2019

(P. Zanuttigh – ICT+Physics of Data)

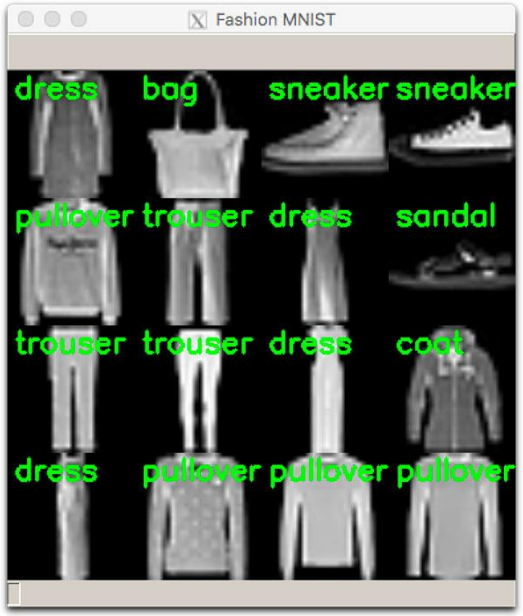
2 Notebooks

- Regularized classification on student alcohol dataset
- SVM classification of clothes images

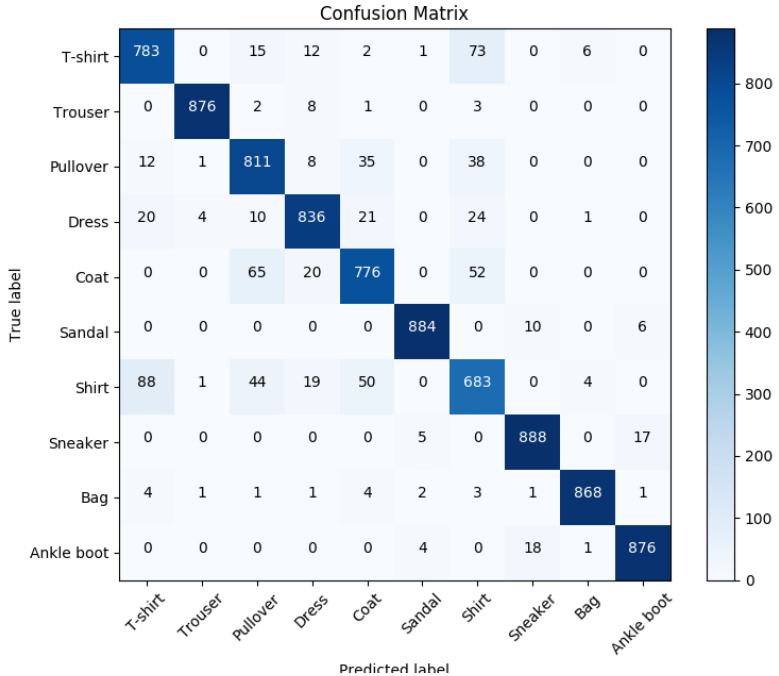
Regularized Classification

- Estimate if a student drinks based on school and personal data
- Perform logistic regression with and without regularization
- Estimate the optimal λ parameter with cross validation
- See the impact of regularization
- Perform a new test with more training data
- Overall test accuracy is relatively low (65-70%) but the impact of regularization can be seen

Classification of Clothes Images



- Dataset of small pictures of clothes: multi-class classification
- Use Support Vector Machines
- Try different Kernels
- Estimate parameters with cross validation
- Visualize the results with confusion matrices



One-hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

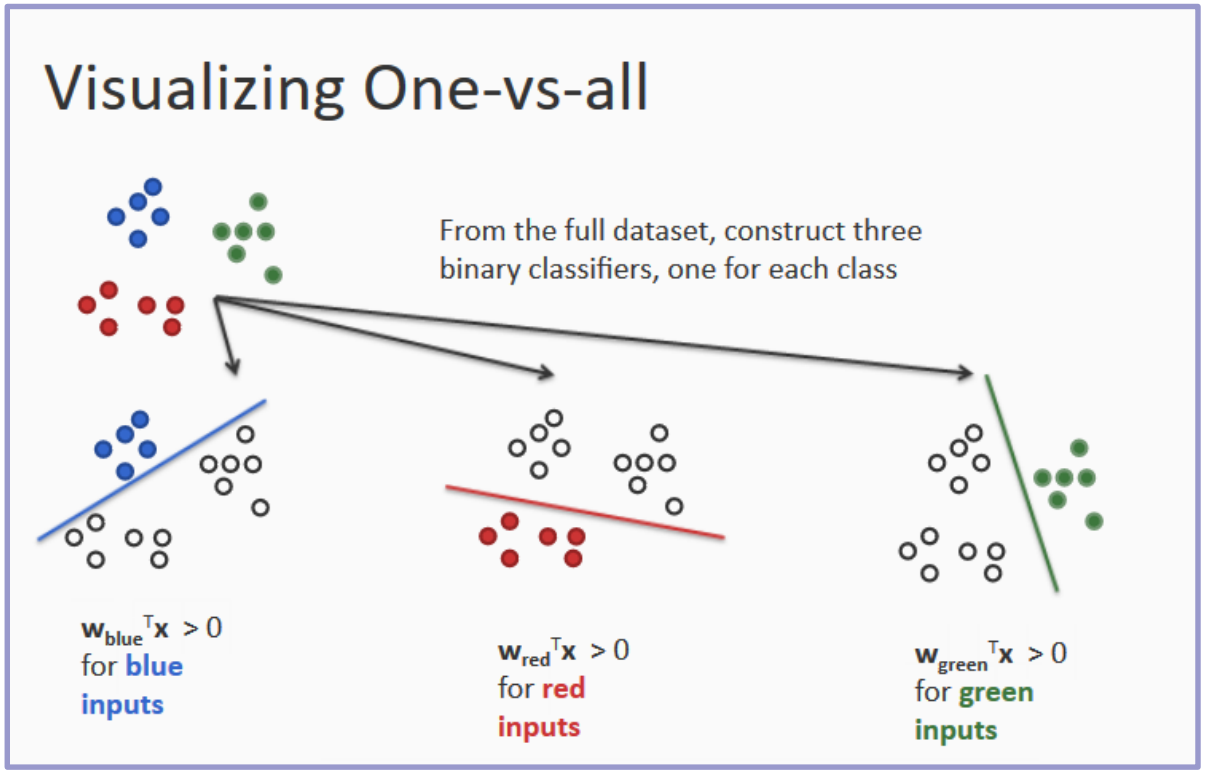
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

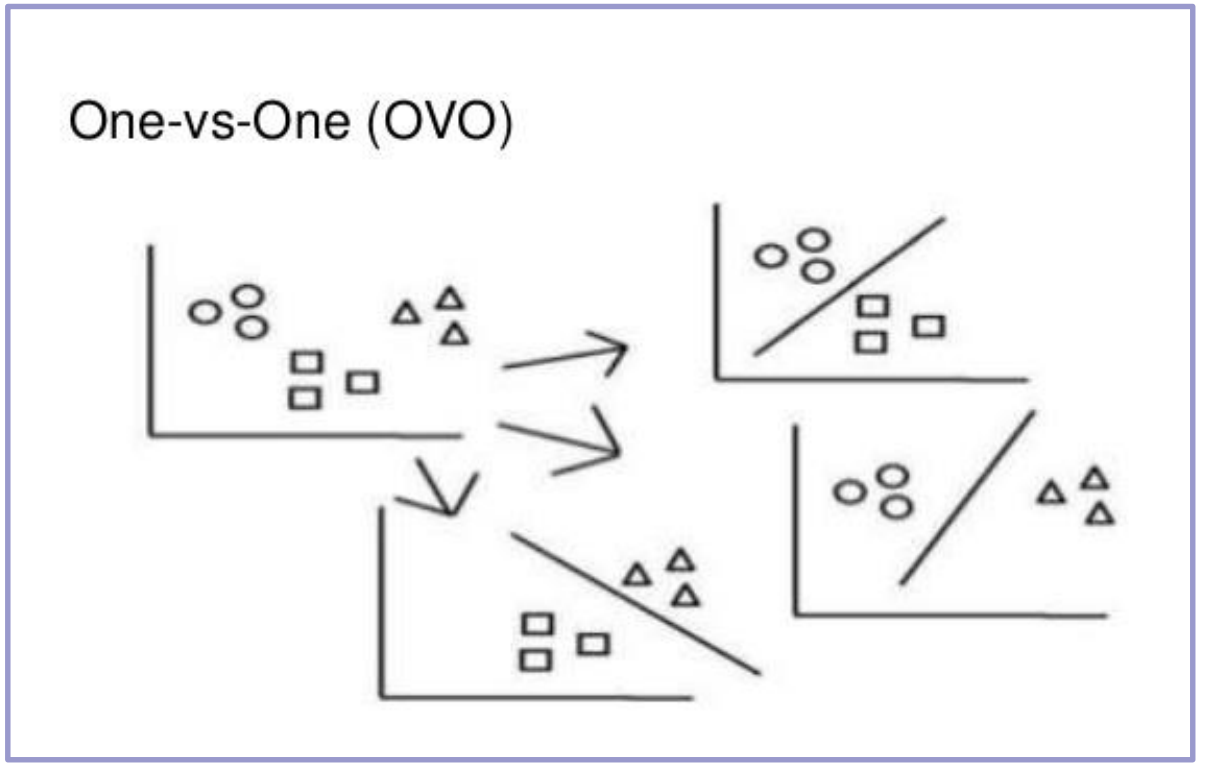
state
NY
WA
CA

AL	…	CA	…	NY	…	WA	…	WY
0	…	0	…	1	…	0	…	0
0	…	0	…	0	…	1	…	0
0	…	1	…	0	…	0	…	0

- One variable for each class (=1 if sample in class, 0 otherwise)
- Avoid having some classes "closer" to others
- Increases data dimensionality

Multi-class classification





- Classify each class vs the union of the others
- For each sample select the class with highest classification score, i.e. $\text{argmax} < \boldsymbol{w_i}, \boldsymbol{x} >$
- Requires $n_{classes}$ comparisons

- Classify each class vs each other class
- For each sample select the class that has "won" the largest number of classifications
- Requires $(\frac{n_{classes}-1}{2})^2$ comparisons
- Used by sklearn