Recap PCA, LDA and pipelines

Plan and recap of last lecture

see Ch. 05 in book “Python Machine Learning” by Raschka & Mirjalili

Agenda

- Recap
  - PCA and LDA
  - Pipelines in scikit-learn
- N-fold cross-validation (CV)
- Interpreting validation curves
- Scikit-learns *GridSearchCV*, and *RandomizedSearchCV* methods for hyperparameter optimization
- Evaluation metrics (if we have time)

Reminder

- Remember that the deadline for CA3 is **tonight at 23:59**

Principal component analysis versus linear discriminant analysis

PCA

Mean centered
data (or standard scaled)
data matrix

$\boldsymbol{X}$

EVD of Cov($\boldsymbol{X}$) to find the tranformation matrix

$$\boldsymbol{x}' = \boldsymbol{x}\boldsymbol{W}$$

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{W}$$

LDA

$$\boldsymbol{m}_i = \frac{1}{n_i}\sum_{\boldsymbol{x}\in D_i}^{c}\boldsymbol{x}_m \qquad \boldsymbol{S}_i = \sum_{\boldsymbol{x}\in D_i}^{c}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T$$

$$\boldsymbol{S}_W = \sum_{i=1}^{c}\boldsymbol{S}_i \qquad \boldsymbol{S}_B = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T$$

In case of unbalanced classes

$$\Sigma_i = \frac{1}{n_i}\boldsymbol{S}_W = \frac{1}{n_i}\sum_{\boldsymbol{x}\in D_i}^{c}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T$$

$\boldsymbol{S}_w^{-1}\boldsymbol{S}_B$:

Limits number of features vs samples,

could combine with PCA.

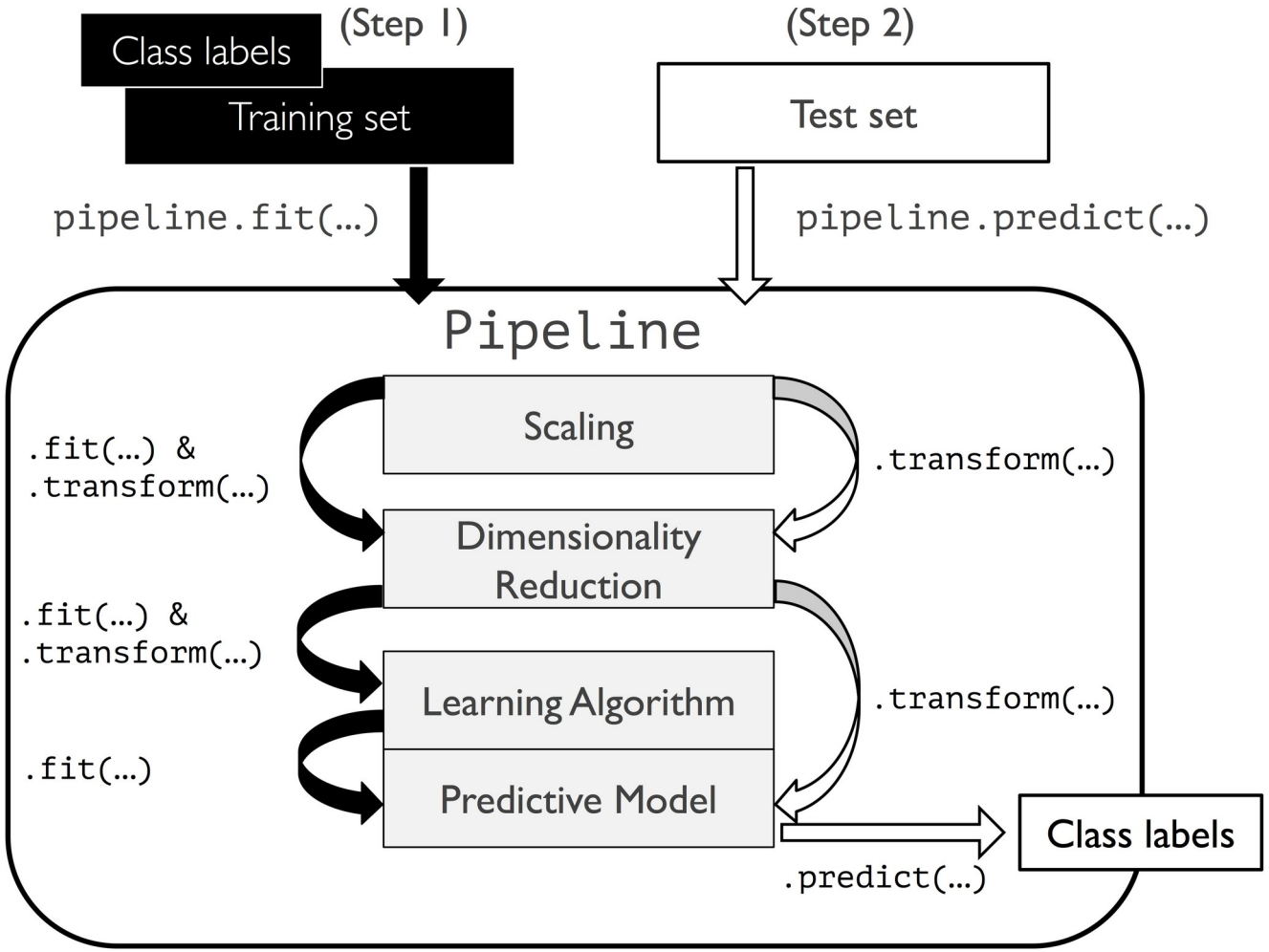
Eigenvaluedecomp. of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_B$: to find tranformation matrix

Max number of LD's: c - 1

$$\boldsymbol{x}' = \boldsymbol{x}\boldsymbol{W}$$

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{W}$$

Pipelines


Class labels
(Step 1)
Training set
(Step 2)
Test set
pipeline.fit(...)
pipeline.predict(...)
Pipeline
Scaling
.fit(...) &
.transform(...)
.transform(...)
Dimensionality
Reduction
.fit(...) &
.transform(...)
Learning Algorithm
.transform(...)
.fit(...)
Predictive Model
Class labels
.predict(...)

Thank you for listening