Applied Machine Learning

Introduction to Ensembles

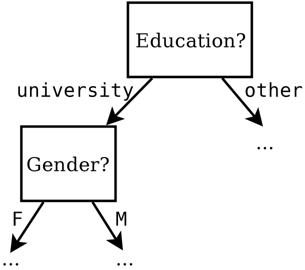
UNIVERSITY OF
GOTHENBURG

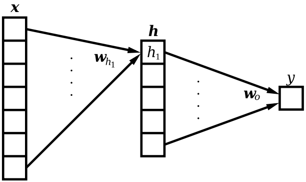
CHALMERS

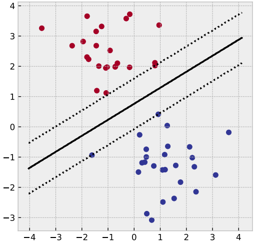
Richard Johansson

richard.johansson@cse.gu.se

why not use more than one?







ensembles

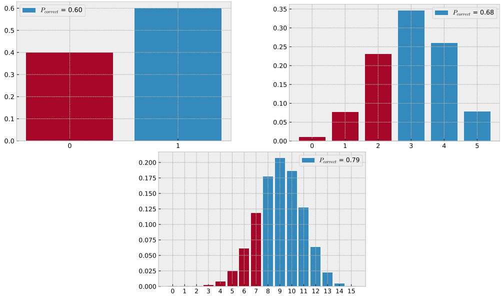
- **ensembles** are machine learning models (classifiers, regressors, rankers, . . .) built by combining several models

[source]

- ensembles often have excellent performance
- when do we expect them to work?

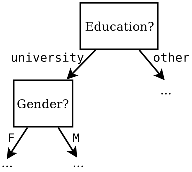
motivation in terms of probabilities

- if we have *n* classifiers whose errors are **independent**, and an accuracy of 0.6, what's the probability that the majority of them are correct?

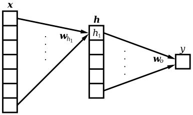


- if the classifiers are **diverse**, it is more likely that they can complement each other how do we implement ensembles?

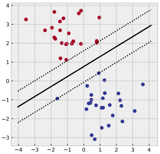
voting


Education?
university
other
...
Gender?
F
M
...
...

⇒ +1

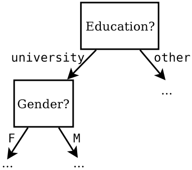

x
h
h_1
w_{h_1}
w_o
y

⇒ -1

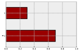

4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

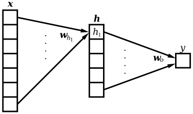
⇒ +1

averaging

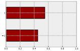

Education?
university
other
...
Gender?
F
M
...
...

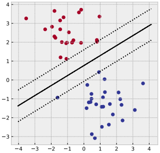
⇒



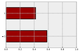

$\boldsymbol{x}$
$\boldsymbol{h}$
h_1
$\boldsymbol{w}_{h_1}$
$\boldsymbol{w}_o$
y

⇒

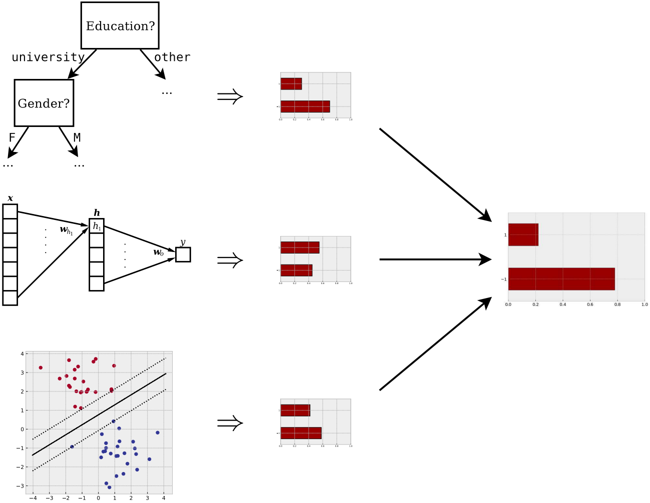



4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

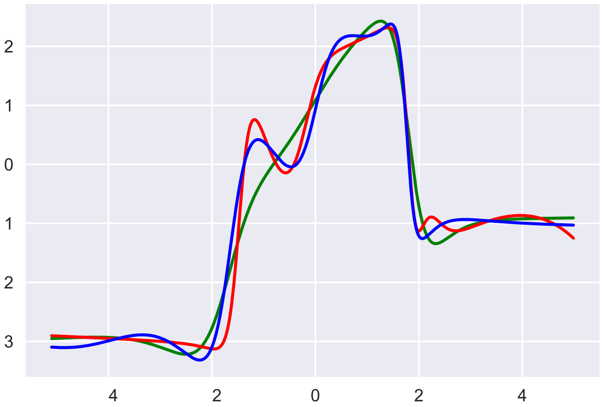
⇒



stacking


Education?
university
other
Gender?
F
M
...
x
h
h_1
w_{h_1}
w_o
y
⇒
1
−1
0.0
0.2
0.4
0.6
0.8
1.0
4
3
2
1
0
−1
−2
−3
−4

what about regression?



how can we create ensembles?

training ensembles: main idea

- we discussed that models in an ensemble should be **diverse**
- is there a way to train a set of models in a way that makes them diverse?

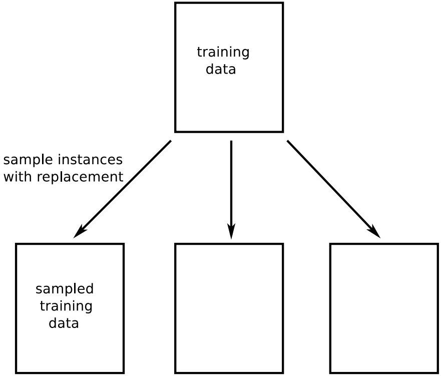
training ensembles: main idea

- we discussed that models in an ensemble should be **diverse**
- is there a way to train a set of models in a way that makes them diverse?
- one idea: make several copies of the training set, where each copy has been modified randomly

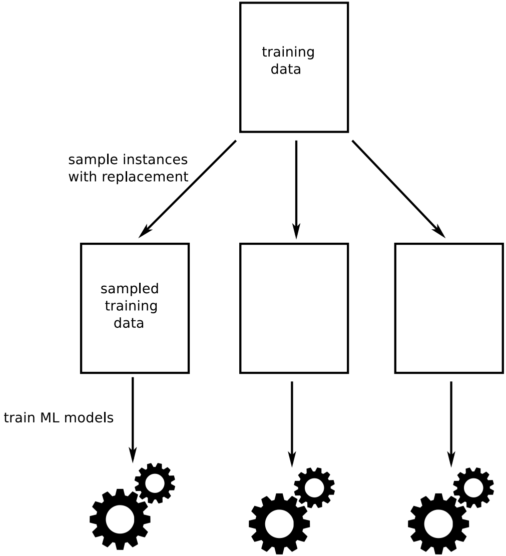
bagging: bootstrap aggregating (Breiman, 1996)

training data

bagging: bootstrap aggregating (Breiman, 1996)

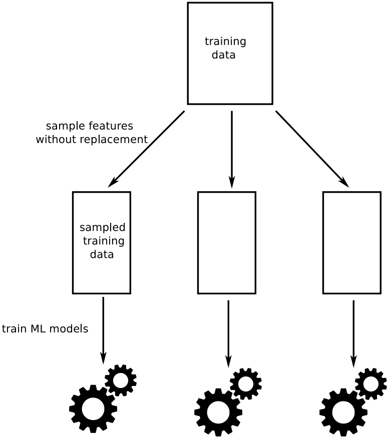


bagging: bootstrap aggregating (Breiman, 1996)



training on random subsets of features

- another spin on the same idea: **feature bagging** or **random subspace learning** (Ho, 1998)
- this procedure creates new training sets by picking random subsets of **features**

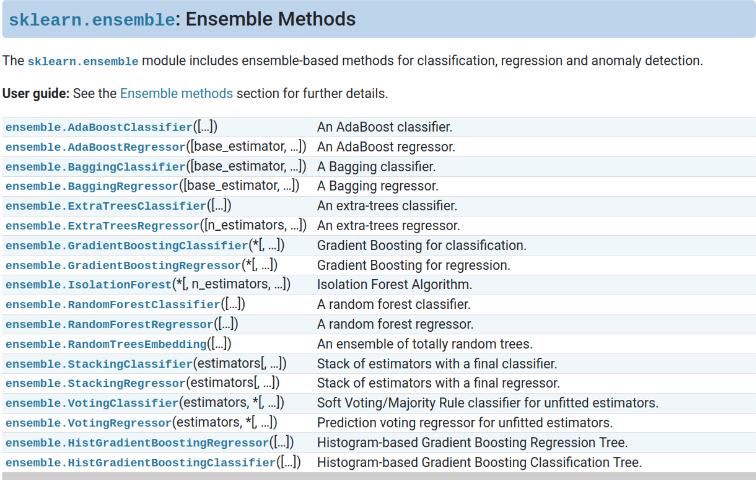


in scikit-learn

```
ensemble = [
            (’lr’, LogisticRegression()),
            (’dt’, DecisionTreeClassifier(max_depth=5)),
            (’svc’, LinearSVC()),
            (’lr1’, LogisticRegression(penalty=’l1’)),
            (’mlp’, MLPClassifier(hidden_layer_sizes=(8),
                                  max_iter=10000))
           ]

pipeline = make_pipeline(
    DictVectorizer(),
    StandardScaler(with_mean=False),
    VotingClassifier(ensemble)
)
```

various types of ensembles in scikit-learn



sklearn.ensemble: Ensemble Methods

The sklearn.ensemble module includes ensemble-based methods for classification, regression and anomaly detection.

User guide: See the Ensemble methods section for further details.

ensemble.AdaBoostClassifier([...])	An AdaBoost classifier.
ensemble.AdaBoostRegressor([base_estimator, ...])	An AdaBoost regressor.
ensemble.BaggingClassifier([base_estimator, ...])	A Bagging classifier.
ensemble.BaggingRegressor([base_estimator, ...])	A Bagging regressor.
ensemble.ExtraTreesClassifier([...])	An extra-trees classifier.
ensemble.ExtraTreesRegressor([n_estimators, ...])	An extra-trees regressor.
ensemble.GradientBoostingClassifier(*[, ...])	Gradient Boosting for classification.
ensemble.GradientBoostingRegressor(*[, ...])	Gradient Boosting for regression.
ensemble.IsolationForest(*[, n_estimators, ...])	Isolation Forest Algorithm.
ensemble.RandomForestClassifier([...])	A random forest classifier.
ensemble.RandomForestRegressor([...])	A random forest regressor.
ensemble.RandomTreesEmbedding([...])	An ensemble of totally random trees.
ensemble.StackingClassifier(estimators[, ...])	Stack of estimators with a final classifier.
ensemble.StackingRegressor(estimators[, ...])	Stack of estimators with a final regressor.
ensemble.VotingClassifier(estimators, *[, ...])	Soft Voting/Majority Rule classifier for unfitted estimators.
ensemble.VotingRegressor(estimators, *[, ...])	Prediction voting regressor for unfitted estimators.
ensemble.HistGradientBoostingRegressor([...])	Histogram-based Gradient Boosting Regression Tree.
ensemble.HistGradientBoostingClassifier([...])	Histogram-based Gradient Boosting Classification Tree.

references

L. Breiman. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

T. K. Ho. 1998. The random subspace method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20(8):832–844.