Permutation Tests for Studying Classifier Performance

Markus Ojala Gemma C. Garriga

Helsinki Institute for Information Technology HIIT
Department of Information and Computer Science
Helsinki University of Technology, Finland

Introduction

Classifier trained on some labeled data:

- Classifier accuracy = quality?
- Class structure = interesting or random artefact?
- Dependency between features?

Two tests for classifiers:

PermClass: Did classifier learn a true class structure?

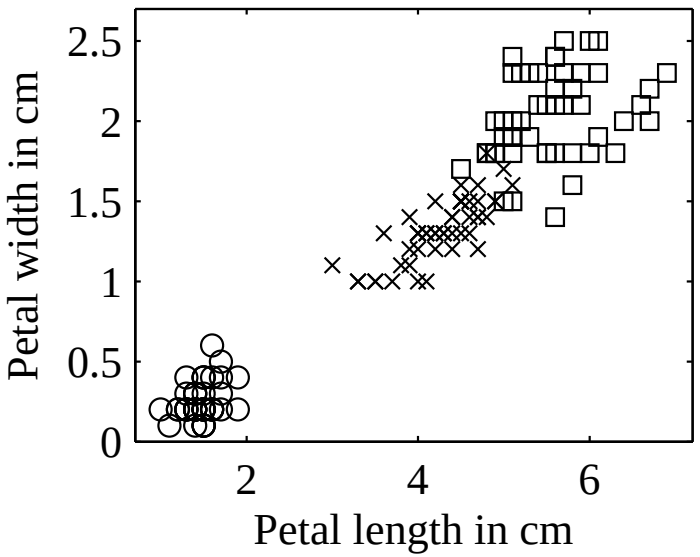
PermFeats: Did classifier exploit dependency between features?

Permutation test: **PermClass**

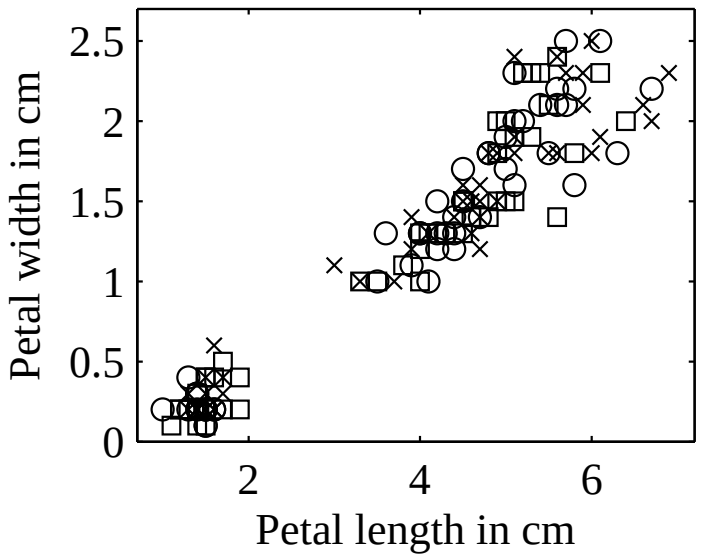
PermClass: Classifier learns a true class structure?

- Null hypothesis: Data and class labels are independent
- Randomization: Permute class labels

Original Iris dataset



PermClass: One sample

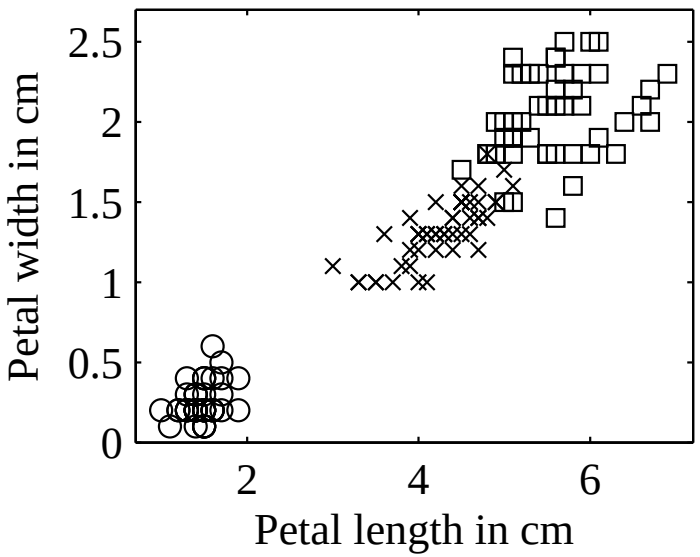


Permutation test: **PermFeats**

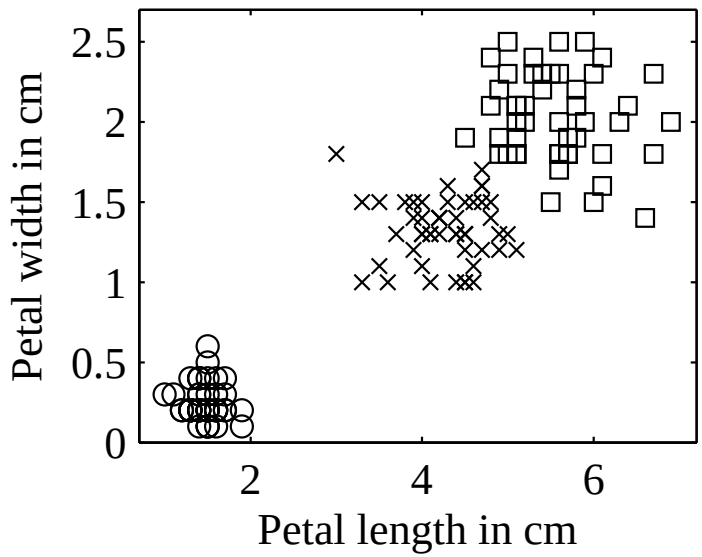
PermFeats: Classifier exploits dependency between features?

- Null hypothesis: Features are mutually independent
- Randomization: Permute features inside each class

Original Iris dataset



PermFeats: One sample



Randomization approach

Setting

- Labeled data $D = \{(X_i, y_i)\}$
- Classifier $f : \mathcal{X} \to \mathcal{Y}$ trained on D
- Classification error $e(f, D)$
  - 10-fold cross-validation error

Testing

- Produce k permutations $\hat{D}$ of original data D from one test
- Teach a classifier $\hat{f}$ for each $\hat{D}$ and calculate $e(\hat{f}, \hat{D})$
- Empirical p-value:

$$p = \frac{|\{\hat{D} : e(\hat{f}, \hat{D}) \leq e(f, D)\}| + 1}{k + 1}$$

Example: 1-Nearest Neighbor on two datasets

Dataset D_1								
o	x	x	x	x	x	x	x	+
x	x	o	x	x	x	x	o	+
x	x	x	x	o	o	x	x	+
x	x	x	x	x	x	x	o	+
x	x	x	x	x	x	x	o	+
x	x	x	x	o	x	x	o	+
o	o	o	x	x	o	o	o	−
o	o	o	o	o	o	o	o	−
x	o	x	o	o	o	o	o	−
x	o	x	o	o	x	o	o	−
o	o	o	o	o	o	x	o	−
o	o	o	x	o	o	o	o	−

Dataset D_2								
x	x	x	o	x	x	x	x	+
x	x	x	x	x	x	x	x	+
x	o	x	x	x	x	x	x	+
o	o	o	o	o	o	o	x	+
o	o	o	o	o	x	o	o	+
o	o	o	o	o	o	o	o	+
x	x	x	x	o	o	o	x	−
x	x	o	x	o	o	o	o	−
x	x	x	x	o	o	o	o	−
o	o	o	o	x	x	x	x	−
o	x	o	o	x	x	x	o	−
o	o	o	x	x	x	x	x	−

Test results for 1-NN:

PermClass: Real class structure (0.001)

PermFeats: No feature dependency (0.358)

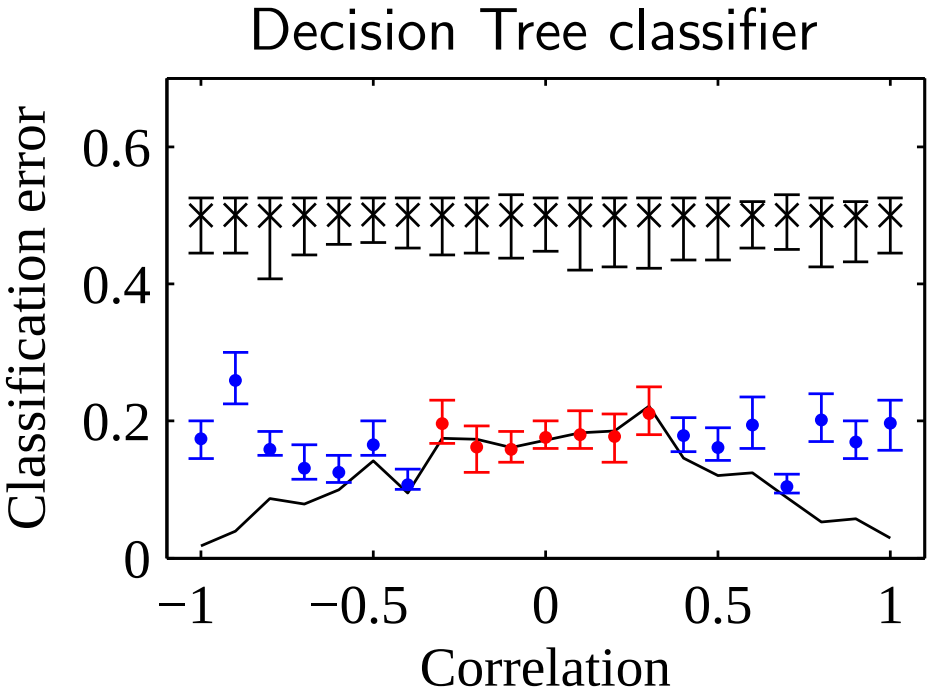
Test results for 1-NN:

Real class structure (0.001)

Feature dependency (0.001)

Generated data: Studying feature dependence

- 21 datasets with varying correlation between 2 features
- In these datasets, correlation increases class separation



× **PermClass**, [1%,99%]

• **PermFeats**, [1%,99%]

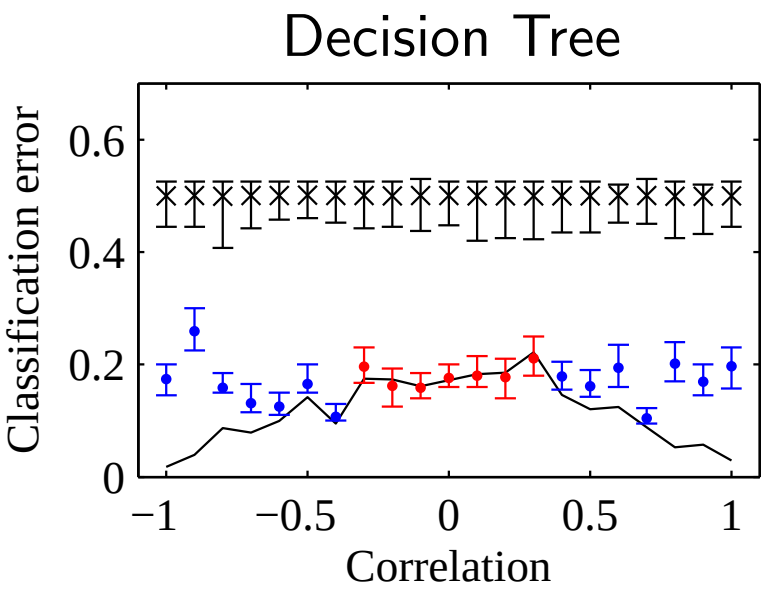
— Original classification error

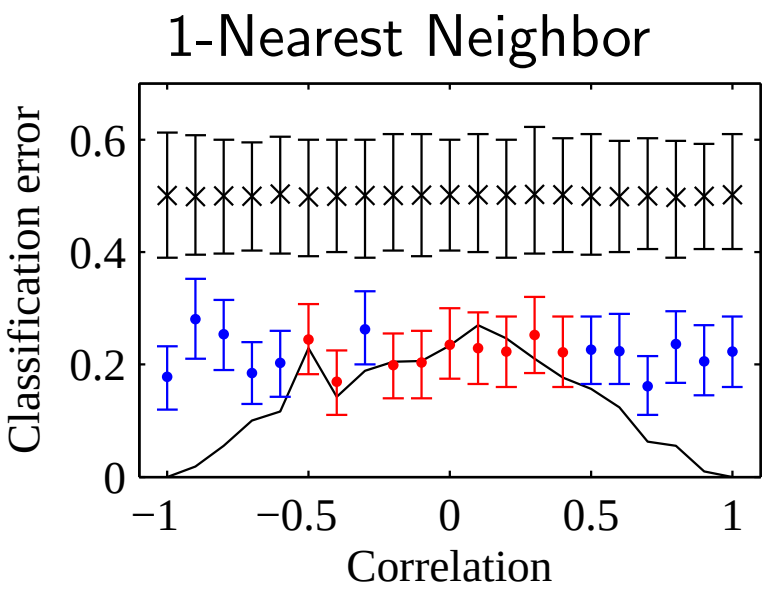
Feature depend. is used

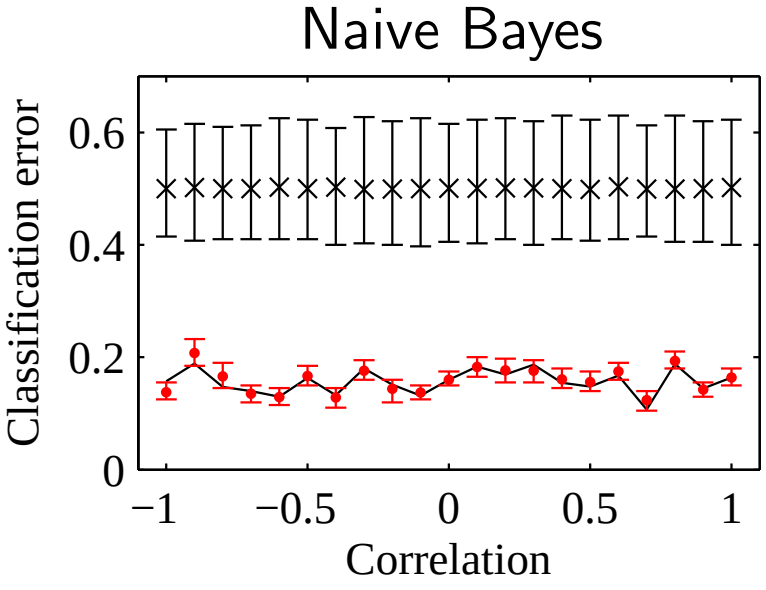
Feature depend. not used

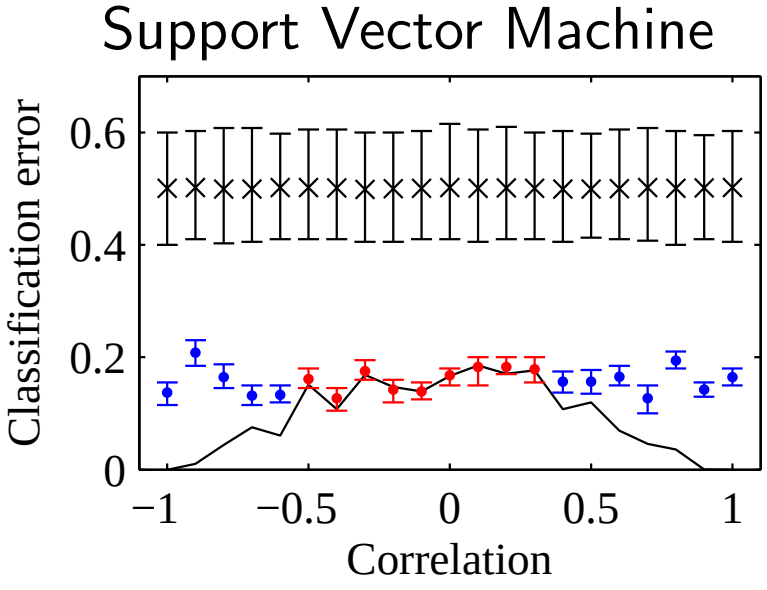
Feature depend. is used

Generated data: Four classifiers









Experiments on real data

Data

- 22 datasets from UCI machine learning repository
- Nominal or/and numeric features

Classifiers (Weka)

- Decision Tree (C4.5)
- 1-Nearest Neighbor
- Naive Bayes
- Support Vector Machine (linear)

P-values: Decision Tree

Dataset	**PermClass**	**PermFeats**
Anneal	0.001	0.001
Autos	0.001	0.001
Balance	0.001	0.001
Mushroom	0.001	0.001
Splice	0.001	0.002
Tic-tac-toe	0.001	0.001
Breast	0.001	0.116
German	0.005	0.666
Iris	0.001	0.765
Pima	0.001	0.642
Promoters	0.002	0.377
Segment	0.001	0.132
Sonar	0.001	0.507
Spect	0.004	0.966
Tumor	0.001	0.138
Votes	0.001	0.791

Feature dependency is used

Feature dependency is not used

P-values: 1-Nearest Neighbor

Dataset	**PermClass**	**PermFeats**
Anneal	0.001	0.001
Autos	0.001	0.001
Balance	0.001	0.001
German	0.001	0.002
Mushroom	0.001	0.001
Sonar	0.001	0.001
Splice	0.001	0.001
Tic-tac-toe	0.001	0.001
Breast	0.007	0.324
Iris	0.001	0.962
Pima	0.001	0.866
Promoters	0.001	0.083
Segment	0.001	0.266
Spect	0.011	0.970
Tumor	0.001	0.860
Votes	0.001	1.000

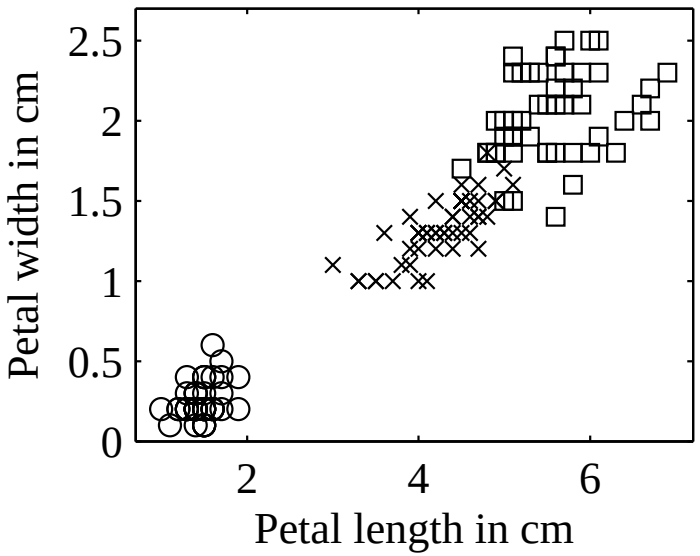
Feature dependency is used

Feature dependency is not used

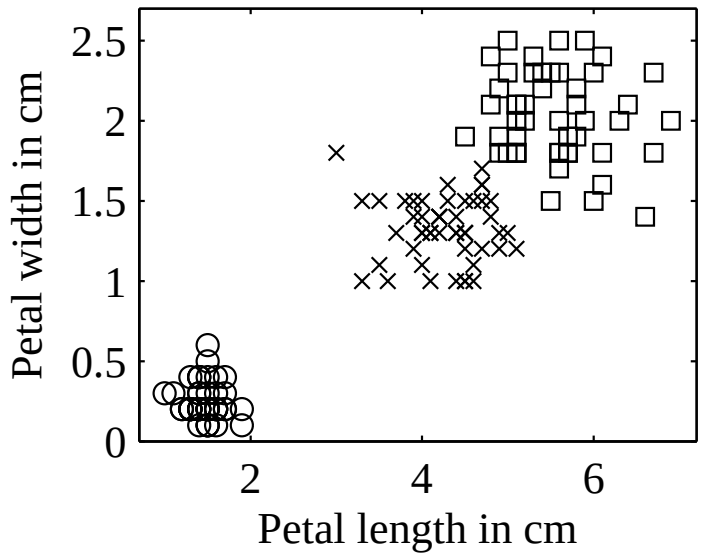
High p-values with **PermFeats**

	Orig.	**PermFeats**	
Dataset	Err.	Av. err.	p-value
Iris	0.05	0.02	0.962

Original Iris dataset



PermFeats: One sample



- Feature dependency can also decrease class separation
- Extreme points are removed in **PermFeats**

Conclusions

- Two permutation tests for studying classifier performance:

PermClass: Did classifier learn a true class structure?

PermFeats: Did classifier exploit dependency between features?

- Experiments showed the usefulness of the test:
  - **PermClass** regards also weak class structure as significant
  - **PermFeats** reveals whether feature dependency is used

Results: 1-Nearest Neighbor

	Orig.	**PermClass**		**PermFeats**	
Dataset	Err.	Av. err.	p-value	Av. err.	p-value
Anneal	0.05	0.40	0.001	0.08	0.001
Autos	0.26	0.77	0.001	0.45	0.001
Balance	0.20	0.56	0.001	0.35	0.001
German	0.28	0.42	0.001	0.33	0.002
Mushroom	0.00	0.50	0.001	0.01	0.001
Sonar	0.13	0.50	0.001	0.27	0.001
Splice	0.24	0.61	0.001	0.30	0.001
Tic-tac-toe	0.21	0.44	0.001	0.38	0.001
Breast	0.31	0.41	0.007	0.32	0.324
Iris	0.05	0.66	0.001	0.02	0.962
Pima	0.29	0.46	0.001	0.27	0.866
Promoters	0.19	0.50	0.001	0.26	0.083
Segment	0.14	0.86	0.001	0.15	0.266
Spect	0.24	0.32	0.011	0.18	0.970
Tumor	0.66	0.88	0.001	0.62	0.860
Votes	0.08	0.47	0.001	0.01	1.000

Feature dependency is used

Feature dependency is not used