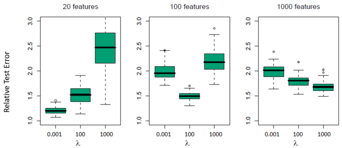
Introduction to Machine Learning

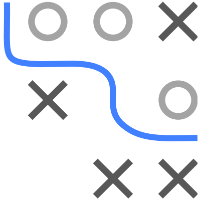
Feature Selection
Feature Selection: Motivating Examples



Learning goals

- Understand the practical importance of feature selection
- Understand that models with integrated selection do not always work
- Know different categories of selection methods

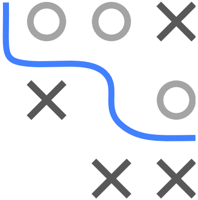
MOTIVATING EXAMPLE 1: REGULARIZATION



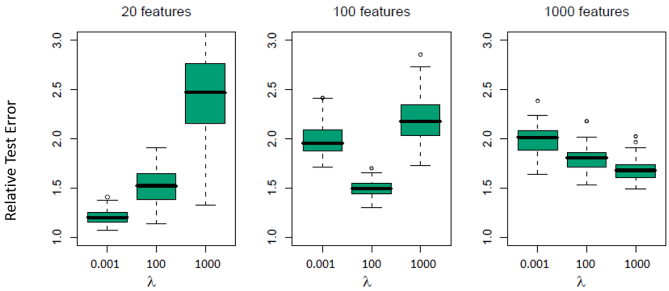
In case of $p \gg n$, overfitting becomes increasingly problematic, as can be shown by the following simulation study:

- For each of 100 simulation iterations:
- Simulate 3 datasets with $n = 100$, $p \in \{20, 100, 1000\}$.
- Features are drawn from a standard Gaussian with pairwise correlation $\rho = 0.2$.
- Target is simulated as $y = \sum_{j=1}^{p} x_j \theta_j + \sigma\varepsilon$, where ε and $\boldsymbol{\theta}$ are both sampled from standard Gaussians, and σ is fixed such that the signal-to-noise ratio is $\text{Var}(\mathbb{E}[y|X])/\sigma^2 = 2$.
- Three ridge regression models with $\lambda \in \{0.001, 100, 1000\}$ are fitted to each simulated dataset.

MOTIVATING EXAMPLE 1: REGULARIZATION / 2



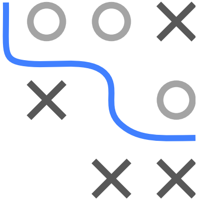
- Boxplots show the relative test error (RTE $=$ test error$/$Bayes error σ^2) over 100 simulations for the different values of p and λ.



- Lowest RTE is obtained at $\lambda = 0.001$ for $p = 20$, at $\lambda = 100$ for $p = 100$, and at $\lambda = 100$ for $p = 100$.
- Optimal amount of regularization increases monotonically in p here.

$\Rightarrow$ High-dimensional settings require more complexity control through regularization or feature selection.

MOTIVATING EX. 2: COMPARISON OF METHODS



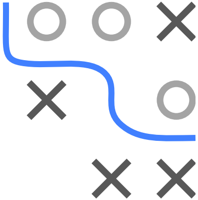
Generalization performance of eight classification methods on micro-array data with $|\mathcal{D}_{\text{train}}| = 144$, $|\mathcal{D}_{\text{test}}| = 54$, $p = 16,063$ genes and a categorical target encoding the type of cancer with 14 classes.

Methods	CV errors (SE) Out of 144	Test errors Out of 54	Number of Genes Used
1. Nearest shrunken centroids	35 (5.0)	17	6,520
2. L_2-penalized discriminant analysis	25 (4.1)	12	16,063
3. Support vector classifier	26 (4.2)	14	16,063
4. Lasso regression (one vs all)	30.7 (1.8)	12.5	1,429
5. k-nearest neighbors	41 (4.6)	26	16,063
6. L_2-penalized multinomial	26 (4.2)	15	16,063
7. L_1-penalized multinomial	17 (2.8)	13	269
8. Elastic-net penalized multinomial	22 (3.7)	11.8	384

Hastie (2009). The Elements of Statistical Learning

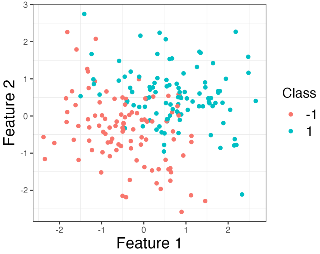
Methods need at least regularization or built-in FS to perform well.
Possible to build good, small models which helps in interpretation

MOTIVATING EX. 3: INTEGRATED SELECTION

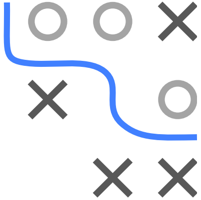


Set-up for simulated micro-array data:

- We generate $n = 200$ samples with $p = 100$ features drawn from a MV Gaussian
- 50% are relevant for the target and 50% have no influence
- Among informative features, 25 are positively and 25 negatively correlated with target, using weights $\{-1, 1\}$
- Target is simulated from Bernoulli distribution using linear predictor as log-odds (linear decision boundary!)



MOTIVATING EX. 3: INTEGRATED SELECTION / 2



- We compare several classifiers regarding their misclassification rate, of which two have integrated FS (rpart and rForest).

- Since we have few observations, we use repeated 10-fold cross-validation with 10 repetitions.

	rpart	lda	logreg	nBayes	knn7	rForest
all feat.	0.44	0.27	0.25	0.32	0.37	0.36
relevant feat.	0.44	0.18	0.19	0.27	0.33	0.30

$\Rightarrow$ Different to Ex. 2, models with integrated FS do not work ideally here. Also, methods with lin. decision boundary are better due to our simulation set-up.

$\Rightarrow$ Performance improves significantly for most methods when only trained on informative features.