UVA CS 4501 - 001 / 6501 – 007 Introduction to Machine Learning and Data Mining

Lecture 22: Feature Selection

Yanjun Qi / Jane, , PhD

University of Virginia
Department of
Computer Science

What we have covered

- ❑ Supervised Regression models
  - Linear regression (LR)
  - LR with non-linear basis functions
  - Locally weighted LR
  - LR with Regularizations
- ❑ Supervised Classification models
  - Support Vector Machine
  - Bayes Classifier
  - Logistic Regression
  - K-nearest Neighbor
  - Random forest / Decision Tree
  - Neural Network (e.g. MLP)

(1) Support Vector Machine

Task → Representation → Score Function → Search/Optimization → Models, Parameters

classification → Kernel Func K(xi, xj) → Margin + Hinge Loss (optional) → QP with Dual form → Dual Weights

$$K(\mathbf{x},z) := \Phi(\mathbf{x})^T \Phi(z)$$

$$w = \sum_i \alpha_i x_i y_i$$

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{p} w_i^2 + C\sum_{i=1}^{n} \varepsilon_i$$

$$\text{subject to } \forall \mathbf{x}_i \in Dtrain : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 - \varepsilon_i$$

$$\underset{k}{\operatorname{argmax}} P(C\_k \mid X) = \underset{k}{\operatorname{argmax}} P(X,C) = \underset{k}{\operatorname{argmax}} P(X \mid C)P(C)$$

(2) Bayes Classifier

Task → Representation → Score Function → Search/Optimization → Models, Parameters

classification → Prob. models p(X|C) → EPE, with Log likelihood(optional) → Many options → Prob. Models' Parameter

$$P(X_1, \cdots, X_p \mid C)$$

Gaussian Naive

$$\hat{P}(X_j \mid C = c_k) = \frac{1}{\sqrt{2\pi}\sigma_{jk}} \exp\left(-\frac{(X_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right)$$

Bernoulli Naïve

$$p(W_i = true \mid c_k) = p_{i,k}$$

Multinomial

$$P(W_1 = n_1, \ldots, W_v = n_v \mid c_k) = \frac{N!}{n_{1k}!\, n_{2k}! \ldots n_{vk}!} \theta_{1k}^{n_{1k}} \theta_{2k}^{n_{2k}} \ldots \theta_{vk}^{n_{vk}}$$

(3) Logistic Regression

Task	Classification
↓	↓
Representation	Log-odds(Y) = linear function of X's
↓	↓
Score Function	EPE, with conditional Log-likelihood
↓	↓
Search/Optimization	Iterative (Newton) method
↓	↓
Models, Parameters	Logistic weights

$$P(c=1|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

(4) K-Nearest Neighbor

Task	Classification
↓	↓
Representation	Local Smoothness
↓	↓
Score Function	NA
↓	↓
Search/Optimization	NA
↓	↓
Models, Parameters	Training Samples

(5) Decision Tree / Random Forest

Task → Representation → Score Function → Search/Optimization → Models, Parameters

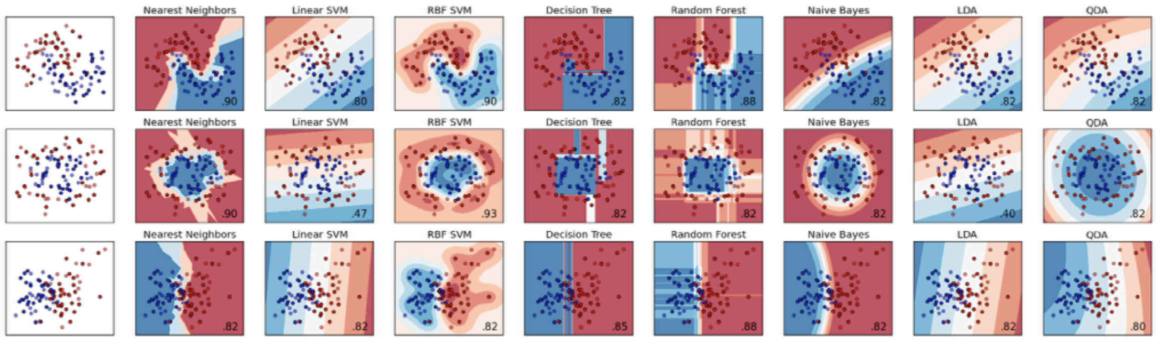
Classification → Partition feature space into set of rectangles → Greedy to find partitions → Split Purity measure / e.g. IG / cross-entropy / Gini / → Tree Model (s), i.e. space partition

Yanjun Qi / UVA CS 4501-01-6501-07

(6) Neural Network

Task → Representation → Score Function → Search/Optimization → Models, Parameters

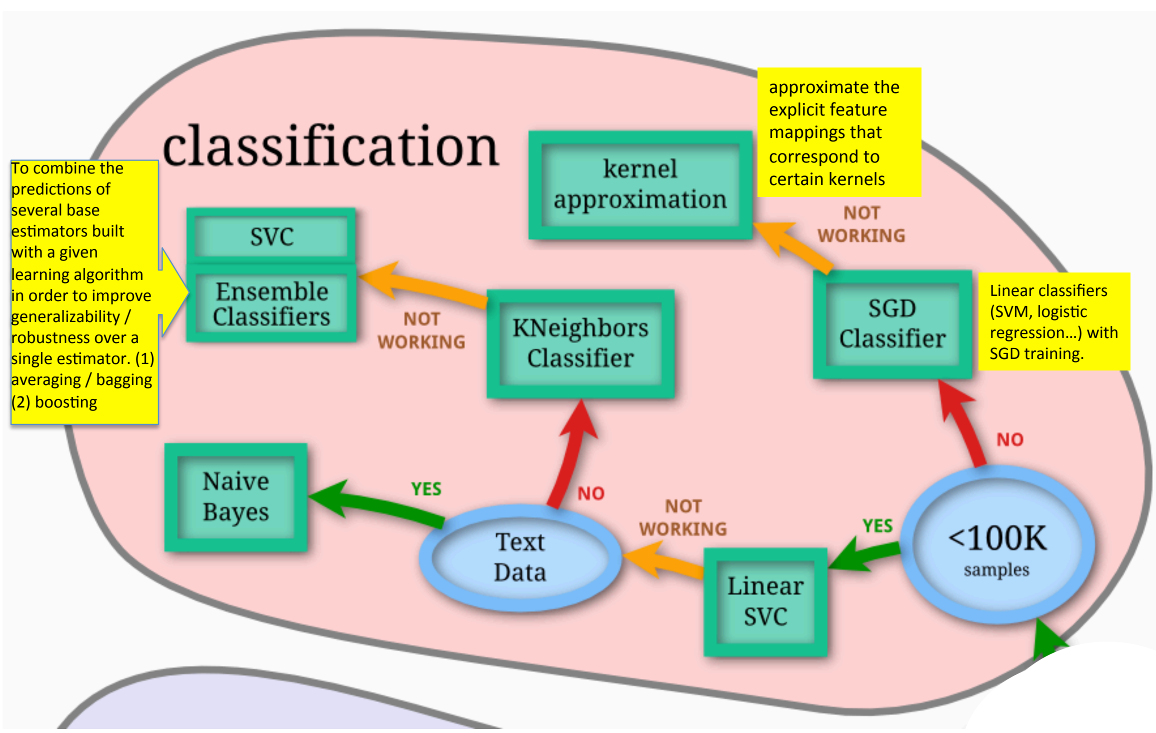
Classification / Regression → Multilayer Network topology → Cross-Entropy / MSE → SGD / Backprop → NN network Weights



- ✓ different assumptions on data
- ✓ different scalability profiles at training time
- ✓ different latencies at prediction time
- ✓ different model sizes (embedability in mobile devices)

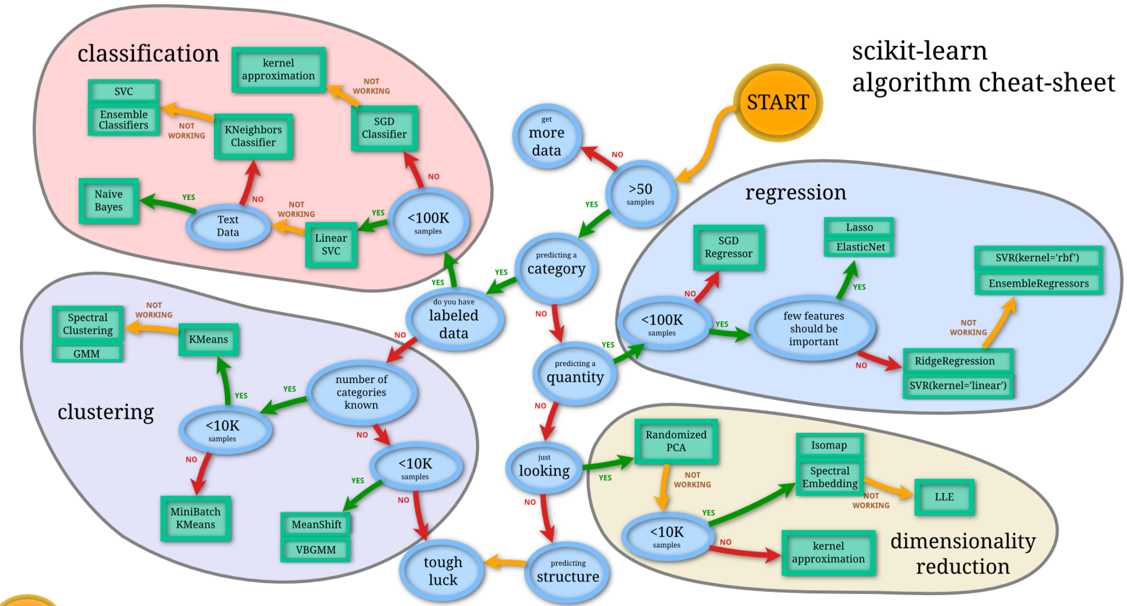
Olivier Grisel's talk

Scikit-learn : Classification



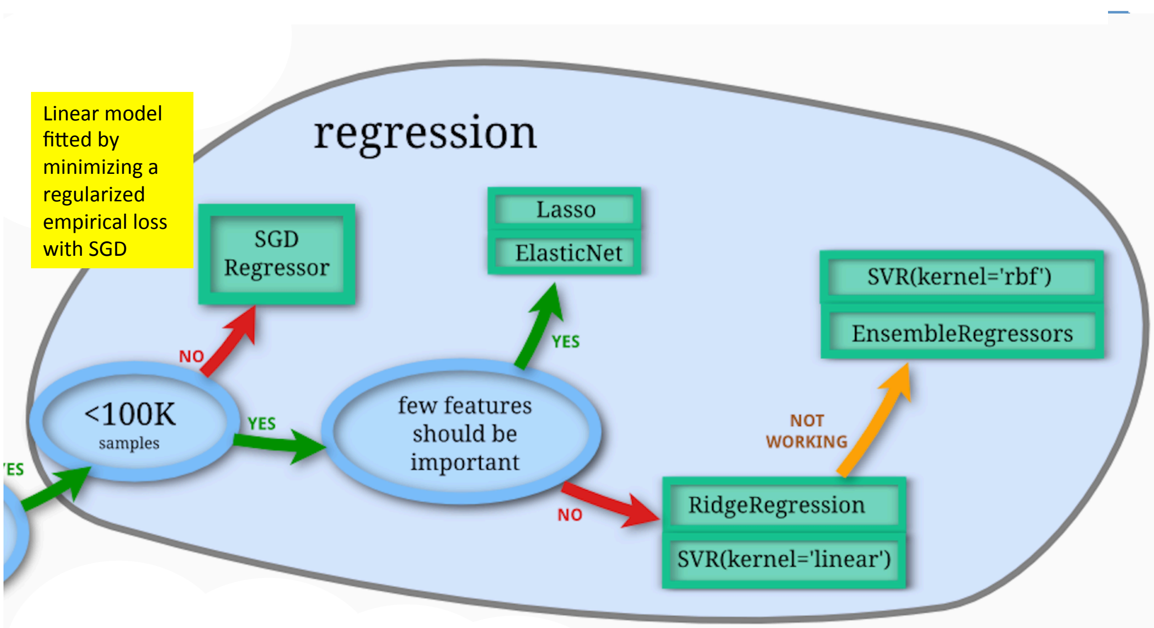
http://scikit-learn.org/stable/tutorial/machine_learning_map/

Scikit-learn algorithm cheat-sheet

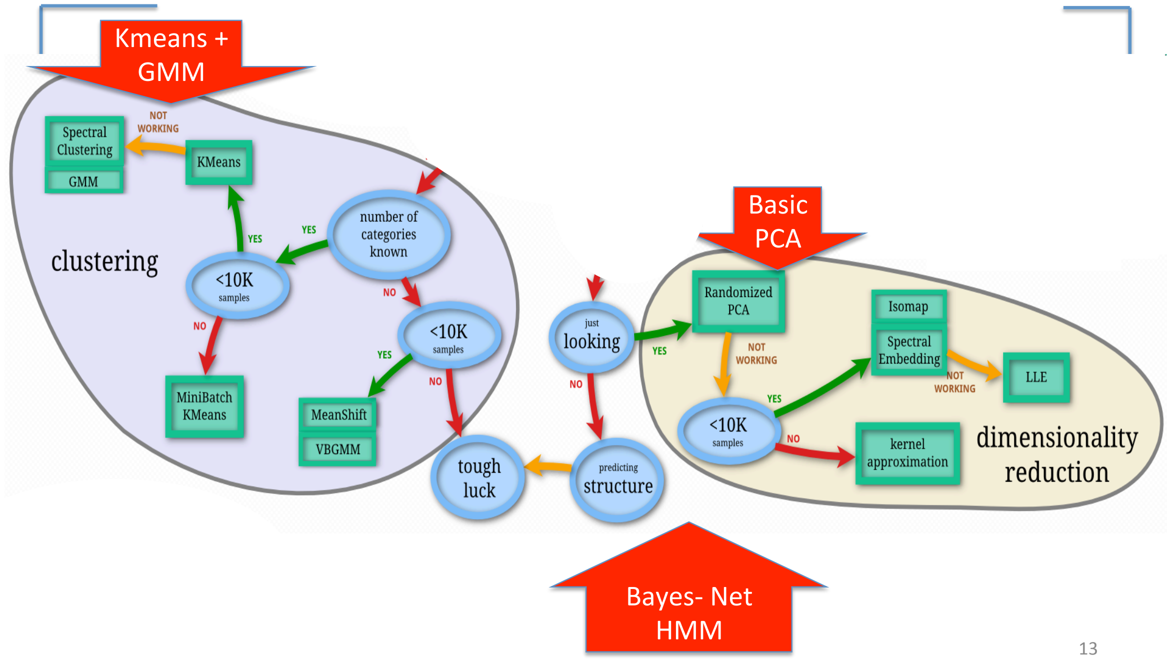


Scikit-learn : Regression

Linear model fitted by minimizing a regularized empirical loss with SGD



next after classification ?

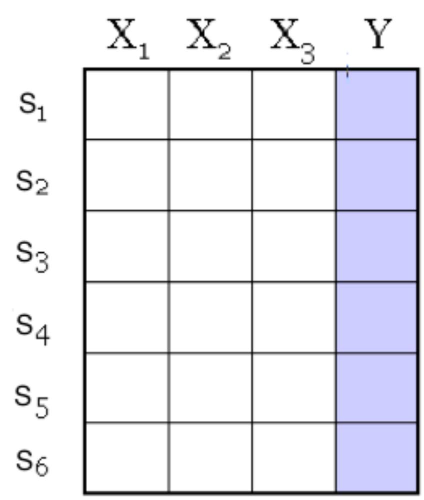


Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
  - ❑ Feature selection
- ❑ Unsupervised models
  - ❑ Dimension Reduction
  - ❑ Clustering
- ❑ Learning theory
- ❑ Graphical models

Today

- Feature Selection (supervised)
  - Filtering approach
  - Wrapper approach
  - Embedded methods



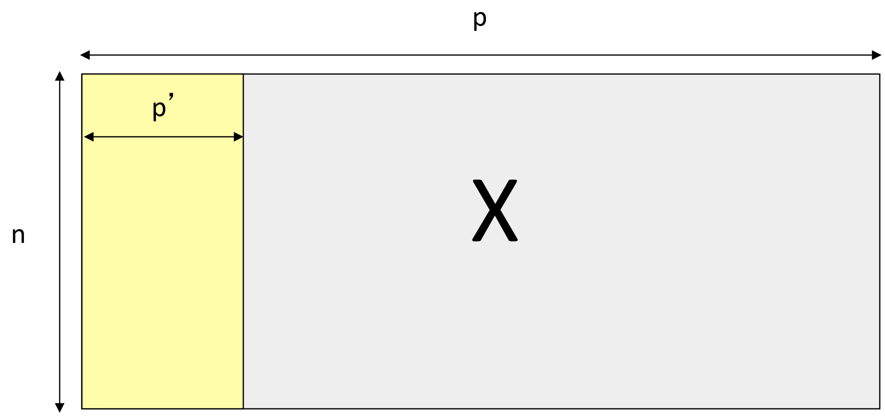
A labeled Dataset

$$f : \boxed{X} \longrightarrow \boxed{Y}$$

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

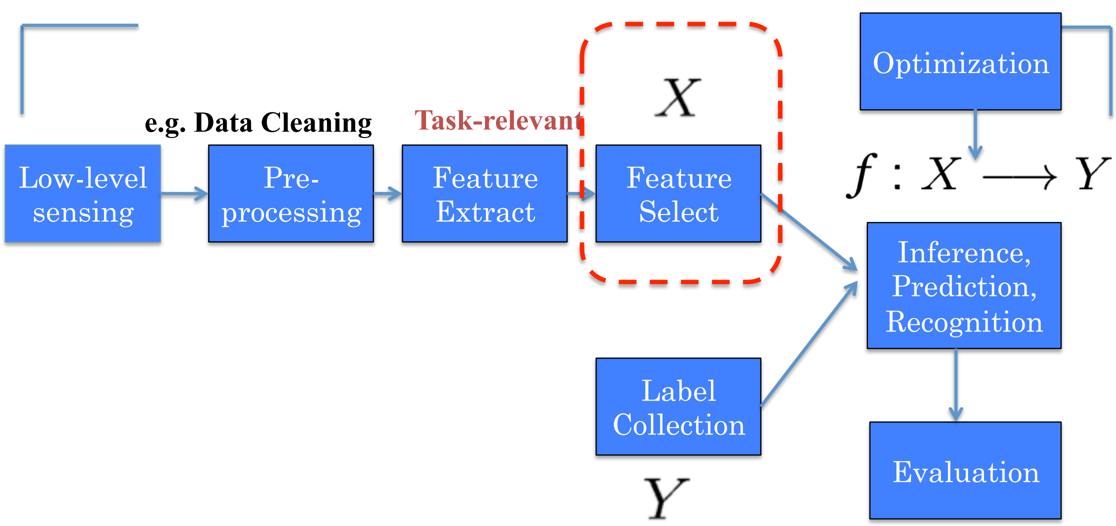
Feature Selection

- **Thousands to millions of low level features**: select the most relevant one to build **better, faster, and easier to understand** learning machines.

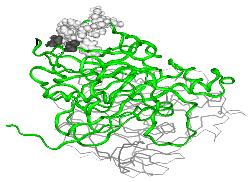


From Dr. **Isabelle Guyon**

A Typical Machine Learning Pipeline



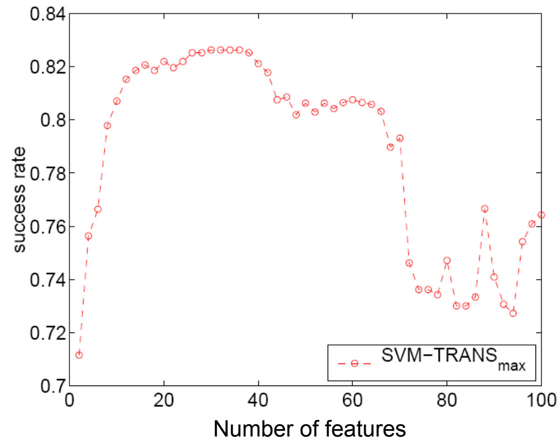
e.g., QSAR: Drug Screening



Binding to Thrombin (DuPont Pharmaceuticals)

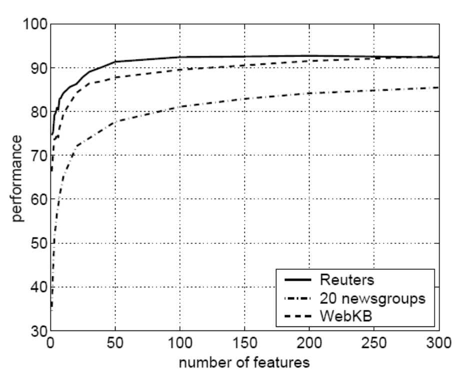
- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.

- **139,351 binary features**, which describe three-dimensional properties of the molecule.



Weston et al, Bioinformatics, 2002

e.g., Text Categorization with feature Filtering



Reuters: 21578 news wire, 114 semantic categories.

20 newsgroups: 19997 articles, 20 categories.

WebKB: 8282 web pages, 7 categories.

Bag-of-words: >100,000 features.

Top 3 words of some output Y categories:

- **Alt.atheism**: atheism, atheists, morality
- **Comp.graphics**: image, jpeg, graphics
- **Sci.space**: space, nasa, orbit
- **Soc.religion.christian**: god, church, sin
- **Talk.politics.mideast**: israel, armenian, turkish
- **Talk.religion.misc**: jesus, god, jehovah

Bekkerman et al, JMLR, 2003

Feature Selection

– Filtering approach:

ranks features or feature subsets independently of the predictor (classifier).

- …using univariate methods: consider one variable at a time
- …using multivariate methods: consider more than one variables at a time

– Wrapper approach:

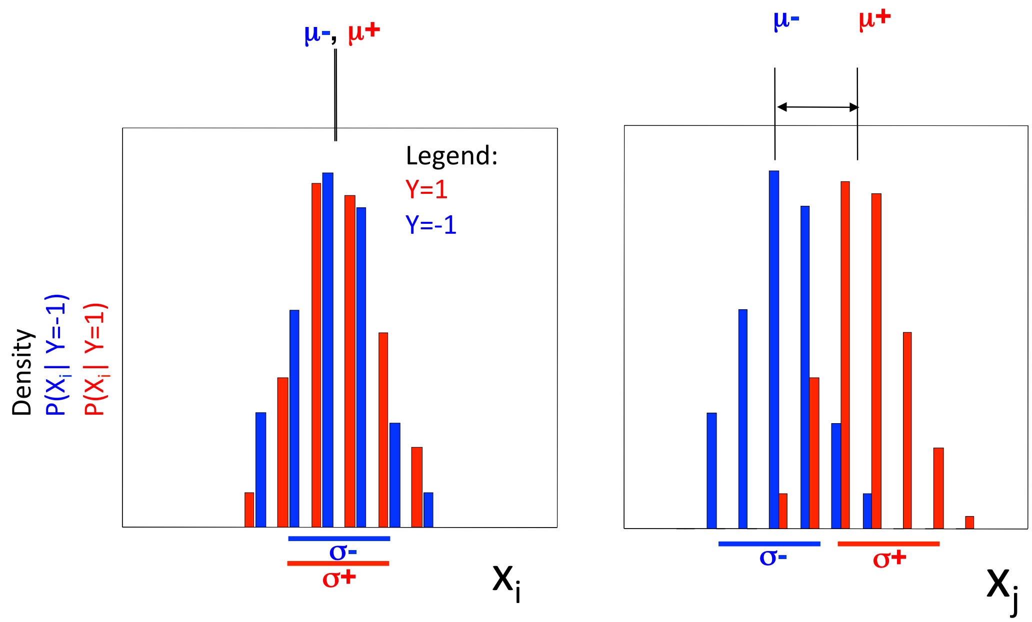
uses a classifier to assess (many) features or feature subsets.

– Embedding approach:

uses a classifier to build a (single) model with a subset of features that are internally selected.

Feature Selection I: univariate filtering approach, e.g. T-test

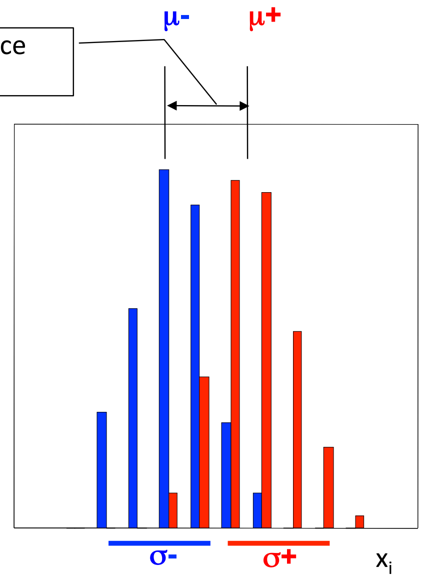
- Issue: determine the relevance of a given single feature.



Feature Selection I: univariate filtering approach , e.g. T-test

T-test

Is this distance significant?



- Normally distributed classes, equal variance σ^2 unknown; estimated from data as σ^2_{within}.
- Null hypothesis H_0: $\mu+ = \mu-$
- T statistic:
  If H_0 is true, then

  $t = (\mu+ - \mu-)/(\sigma_{within}\sqrt{1/m^+ + 1/m^-}) \sim$ Student$(m^+ + m^- - 2$ d.f.)

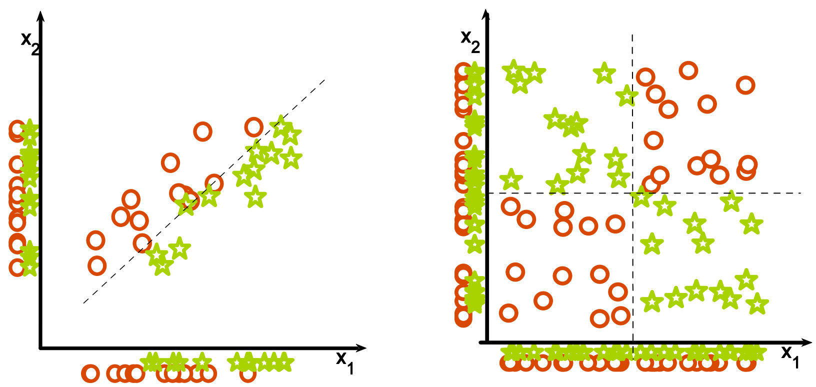
From Dr. **Isabelle Guyon**

Feature Selection I: univariate filtering, (many other criteria)

Method		X			Y			Comments
Name	Formula	B	M	C	B	M	C	
Bayesian accuracy	Eq. 3.1	+	s		+	s		Theoretically the golden standard, rescaled Bayesian relevance Eq. 3.2.
Balanced accuracy	Eq. 3.4	+	s		+	s		Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets.
Bi-normal separation	Eq. 3.5	+	s		+	s		Used in information retrieval.
F-measure	Eq. 3.7	+	s		+	s		Harmonic of recall and precision, popular in information retrieval.
Odds ratio	Eq. 3.6	+	s		+	s		Popular in information retrieval.
Means separation	Eq. 3.10	+	i	+	+			Based on two class means, related to Fisher's criterion.
T-statistics	Eq. 3.11	+	i	+	+			Based also on the means separation.
Pearson correlation	Eq. 3.9	+	i	+	+	i	+	Linear correlation, significance test Eq. 3.12, or a permutation test.
Group correlation	Eq. 3.13	+	i	+	+	i	+	Pearson's coefficient for subset of features.
χ^2	Eq. 3.8	+	s		+	s		Results depend on the number of samples m.
Relief	Eq. 3.15	+	s	+	+	s	+	Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions.
Separability Split Value	Eq. 3.41	+	s	+	+	s		Decision tree index.
Kolmogorov distance	Eq. 3.16	+	s	+	+	s	+	Difference between joint and product probabilities.
Bayesian measure	Eq. 3.16	+	s	+	+	s	+	Same as Vajda entropy Eq. 3.23 and Gini Eq. 3.39.
Kullback-Leibler divergence	Eq. 3.20	+	s	+	+	s	+	Equivalent to mutual information.
Jeffreys-Matusita distance	Eq. 3.22	+	s	+	+	s	+	Rarely used but worth trying.
Value Difference Metric	Eq. 3.22	+	s		+	s		Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations.
Mutual Information	Eq. 3.29	+	s	+	+	s	+	Equivalent to information gain Eq. 3.30.
Information Gain Ratio	Eq. 3.32	+	s	+	+	s	+	Information gain divided by feature entropy, stable evaluation.
Symmetrical Uncertainty	Eq. 3.35	+	s	+	+	s	+	Low bias for multivalued features.
J-measure	Eq. 3.36	+	s	+	+	s	+	Measures information provided by a logical rule.
Weight of evidence	Eq. 3.37	+	s	+	+	s	+	So far rarely used.
MDL	Eq. 3.38	+	s		+	s		Low bias for multivalued features.

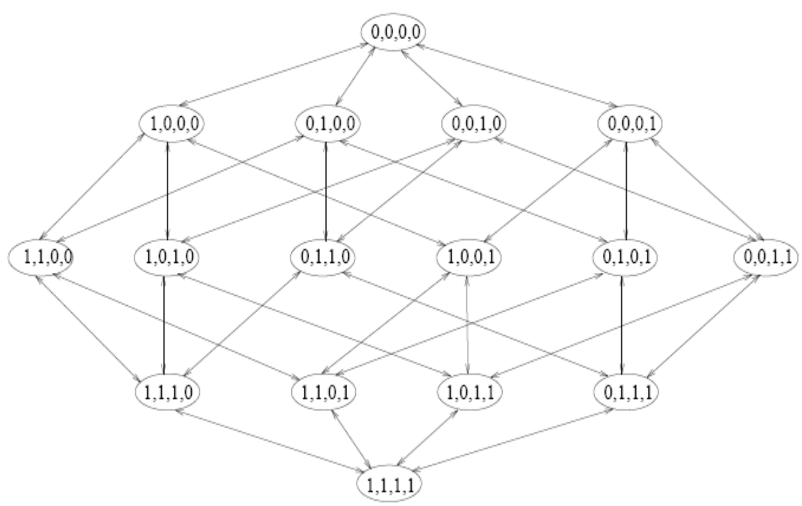
Feature Selection: multivariate approach

Univariate selection may fail



Guyon-Elisseeff, JMLR 2004; Springer 2006

Feature Selection: search strategies



p features, 2^p possible feature subsets!

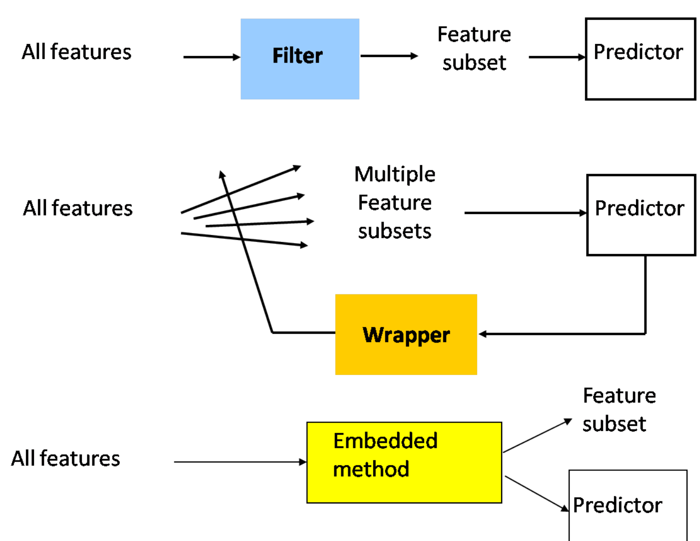
Feature Selection II: search strategies for wrapper approaches

- **Forward selection** or **backward elimination.**
- **Beam search:** keep k best path at each step.
- **GSFS:** generalized sequential forward selection – when (n-k) features are left try all subsets of g features. More trainings at each step, but fewer steps.
- **PTA(l,r):** plus l , take away r – at each step, run SFS l times then SBS r times.
- **Floating search**: One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far.

From Dr. **Isabelle Guyon**

Feature Selection: filters vs. wrappers vs. embedding

- **Main goal**: rank subsets of useful features

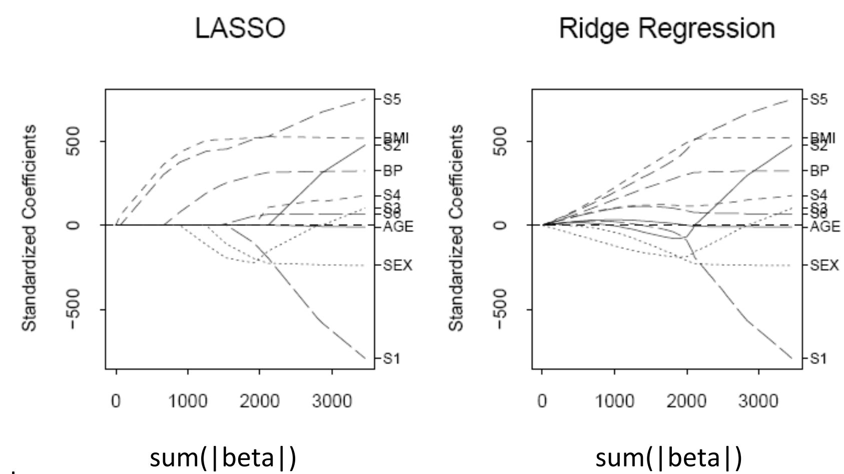


From Dr. **Isabelle Guyon**

Feature Selection III: e.g. **Feature Selection via Embedded Methods: L_1-regularization**

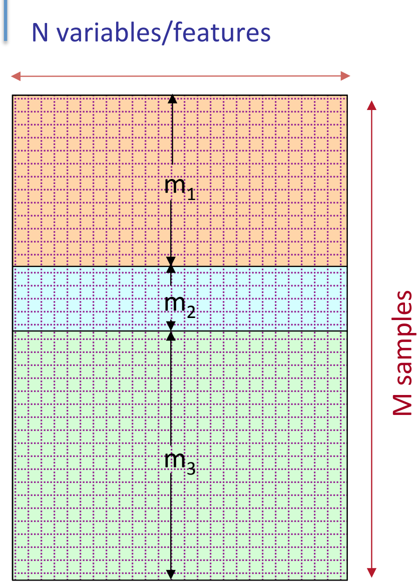
l_1 penalty: $y \sim Model(X\beta) + \lambda \sum |\beta_i|$ (lasso)

l_2 penalty: $y \sim Model(X\beta) + \lambda \sum \beta_i^2$ (ridge regression)



From ESL book

Feature Selection: feature subset assessment (for wrapper approach)



Split data into 3 sets:
training, validation, and test set.

1) For each feature subset, train predictor on training data.
2) Select the feature subset, which performs best on validation data.
   - Repeat and average if you want to reduce variance (cross-validation).
3) Test on test data.

Danger of over-fitting with intensive search!
From Dr. **Isabelle Guyon**

In practice…

- **No method is universally better:**
  - **wide variety of types of variables, data distributions, learning machines, and objectives.**

- **Feature selection is not always necessary to achieve good performance.**

NIPS 2003 and WCCI 2006 challenges : http://clopinet.com/challenges

From Dr. **Isabelle Guyon**

References

- Prof. Andrew Moore's slides
- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.
- Dr. **Isabelle Guyon's feature selection tutorials**