Big Picture

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

March 2nd, 2005

What you have learned thus far

- Learning is function approximation
- Point estimation
- Regression
- Naïve Bayes
- Logistic regression
- Bias-Variance tradeoff
- Neural nets
- Decision trees
- Cross validation
- Boosting
- Instance-based learning
- SVMs
- Kernel trick
- PAC learning
- VC dimension
- Margin bounds
- Mistake bounds

Review material in terms of…

- Types of learning problems
- Hypothesis spaces
- Loss functions
- Optimization algorithms

Text Classification

text $\longrightarrow$ {C, P, U, ...}


the world of
TOTAL

All About The Company
Global Activities
Corporate Structure
TOTAL's Story
Upstream Strategy
Downstream Strategy
Chemicals Strategy
TOTAL Foundation
Homepage

all about the **company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

Company home page

vs

Personal home page

vs

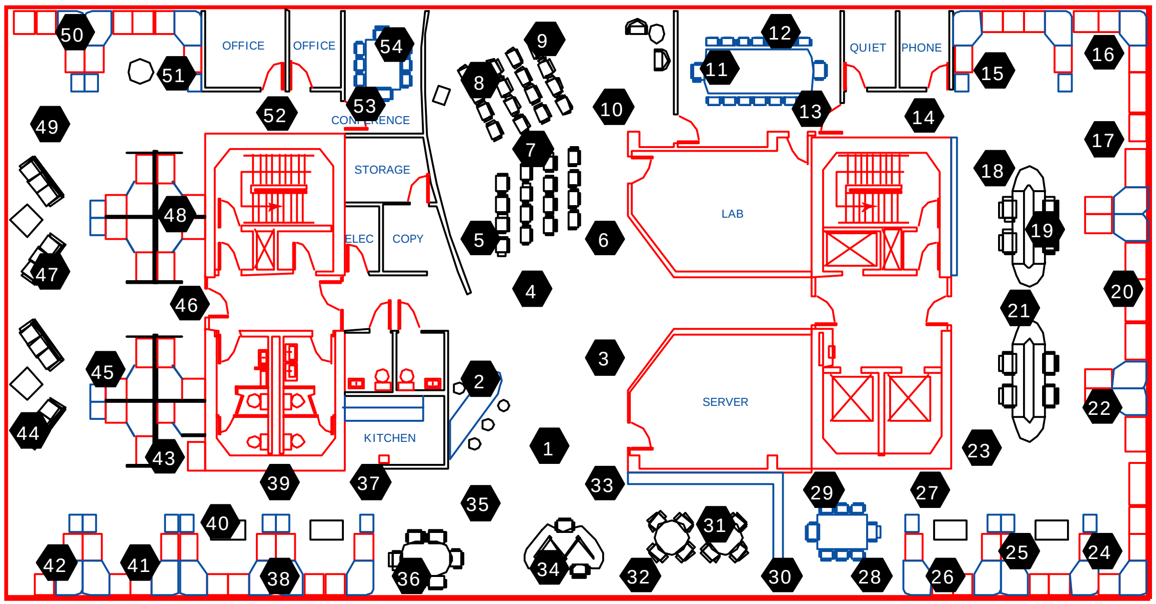
Univeristy home page

vs

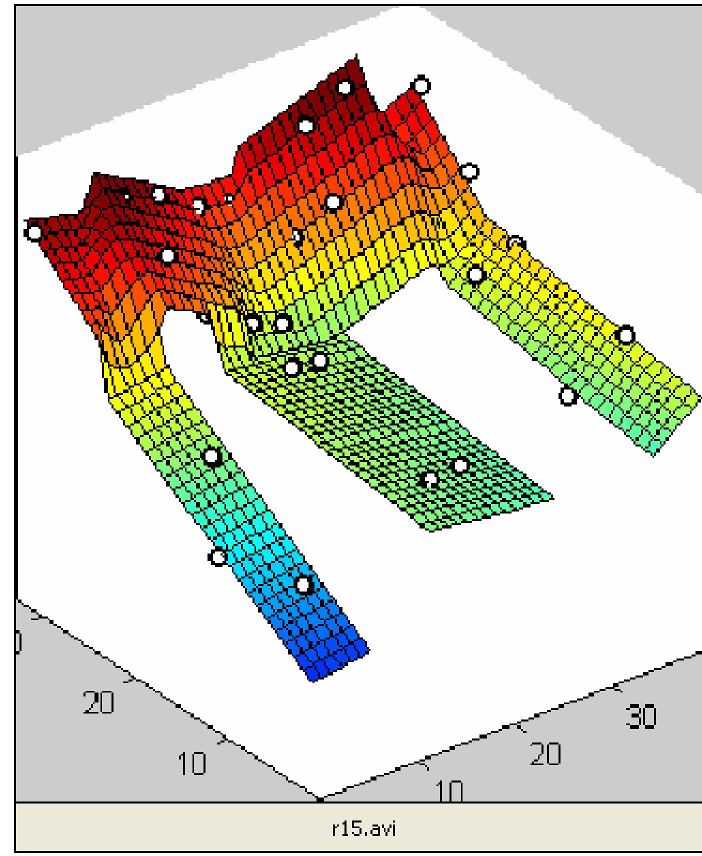
…

Function fitting

x, y, t → temps

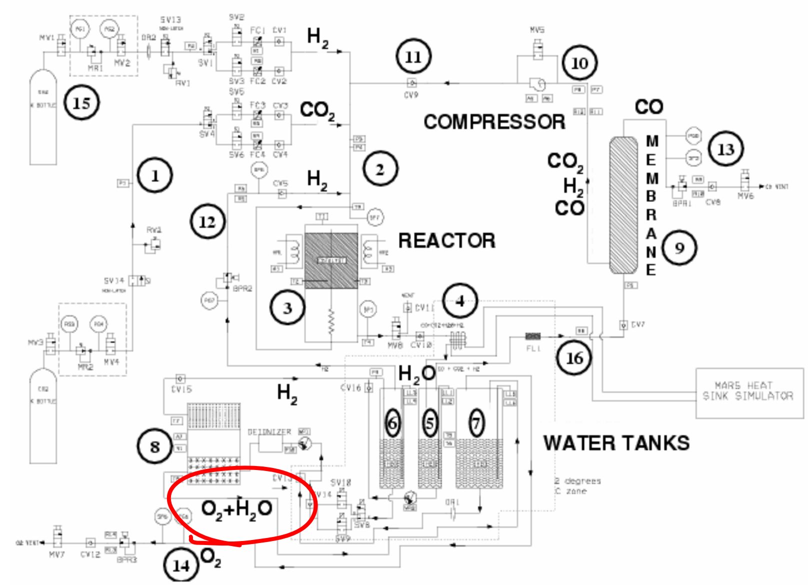


Temperature data



Sensors $\longrightarrow$ probabilities

Monitoring a complex system



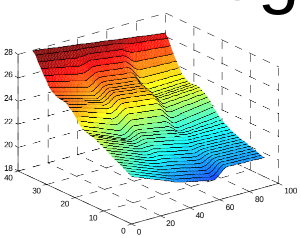
- Reverse water gas shift system (RWGS)
- Learn model of system from data
- Use model to predict behavior and detect faults

Types of learning problems

- Classification

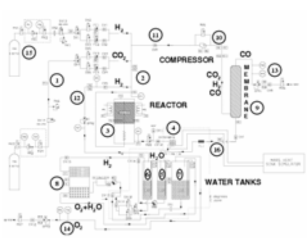
text → {C, P, U, ...}

- Regression



x, y, t → $\mathbb{R}$

- Density estimation



Sensors → [0, 1]

Input – Features

x $\phi(x)$

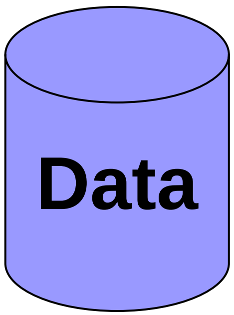
Output?

y: classification: discrete

regression: $\mathbb{R}$

Density Est.: [0, 1]

The learning problem



$\langle x_1,\ldots,x_n,y\rangle$

Learning task

Features/Function approximator

- linear
- NN
- DTs

Loss function

- Squared error
- MLE (log-loss)
- Margin

Optimization algorithm

- Gradient
- QPs

Learned function

$\hat{f}(x)$

unseen
test
data

Comparing learning algorithms

- Hypothesis space
- Loss function
- Optimization algorithm

Naïve Bayes versus Logistic regression

Density estimation: P(Y|x)

Naïve Bayes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(X|Y) = \prod_i P(X_i|Y)$$

strong indep. assumption

loss function:

max P(X,Y)

MLE

Logistic regression

$$P(Y=1|x) = \frac{1}{1 + exp(w_0 + \sum_i w_i x_i)}$$

weaker indep. assumptions

loss: max P(Y|X)

Gradient, MLE

equivalence

indep. Gaussian features

Naïve Bayes versus Logistic regression – Classification as density estimation

$$P(Y|X)$$

- Choose class with highest probability

$$\hat{y} = \arg\max_{y} P(y|x)$$

- In addition to class, we get certainty measure

Logistic regression versus Boosting

Logistic regression

$$P(Y = y_i|\mathbf{x}) = \frac{1}{1 + exp(-y_i(\mathbf{w}.\mathbf{x} + b))}$$

Log-loss

$$\sum_{j=1}^{m} \log\left[1 + exp(-y_i(\mathbf{w}.\mathbf{x}_j + b))\right]$$

→ sign (w.x+b) classifier

→ select features apriori

Global optima

Boosting

optimizing α_t

Classifier

$$sign\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})\right)$$

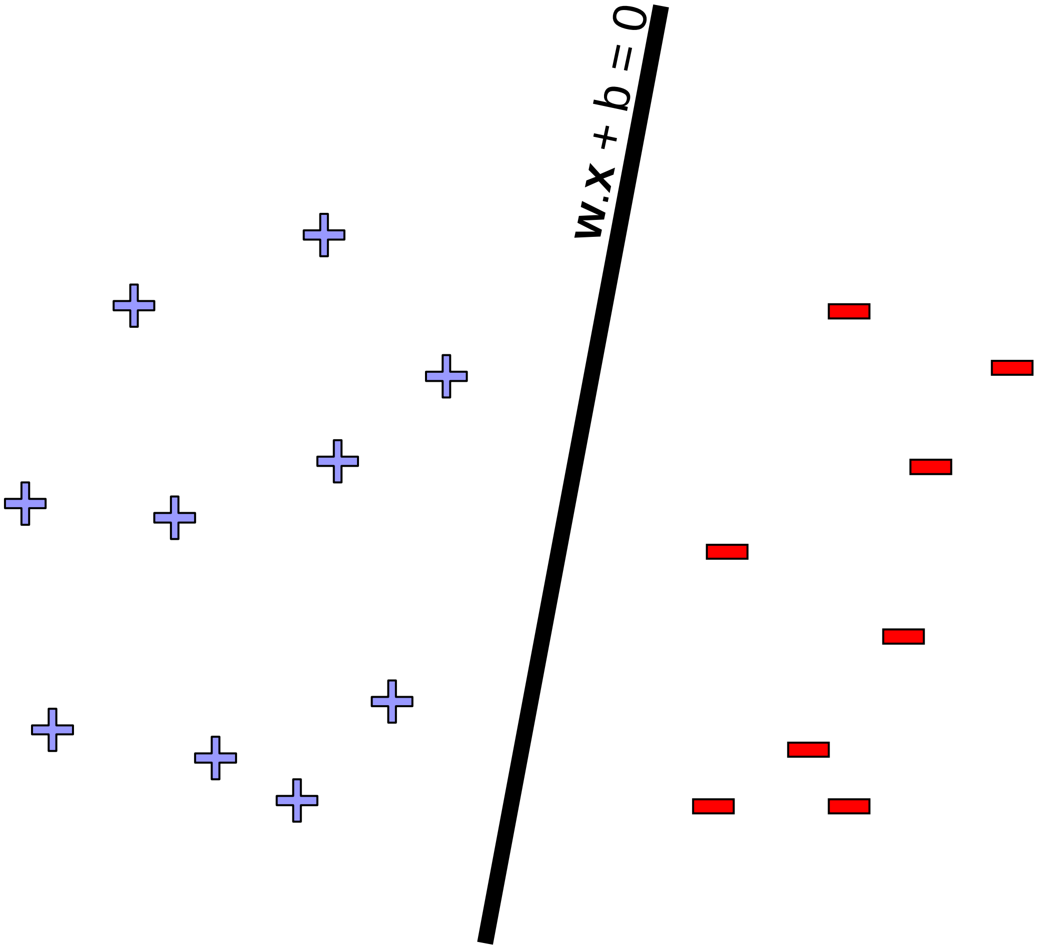
Exponential-loss

$$\frac{1}{m}\sum_{j=1}^{m} \exp\left(-y_j \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x_j})\right)$$

→ weak learners are the features $h_t(x)$

→ sequential optimization

Linear classifiers – Logistic regression versus SVMs



What's the difference between SVMs and Logistic Regression? (Revisited again)

	SVMs	**Logistic Regression**
Loss function	Hinge loss	Log-loss
High dimensional features with kernels	Yes!	Yes!
Solution sparse	Often yes!	Almost always no!
	Classification	$P(Y\|X)$ density estimation

SVMs and instance-based learning

$$\mathbf{w}\cdot\Phi(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x},\mathbf{x}_i)$$

$$b = y_k - \sum_i \alpha_i y_i K(\mathbf{x}_k,\mathbf{x}_i)$$

for any k where $C > \alpha_k > 0$

Classify as

SVMs

$$sign\left(\sum_i \alpha_i y_i K(\mathbf{x},\mathbf{x}_i) + b\right)$$

optimize α_i

Instance based learning

as density estimation

$$P(y\mid\mathbf{x}) = \frac{\sum_i y_i K(\mathbf{x},\mathbf{x}_i)}{\sum_i K(\mathbf{x},\mathbf{x}_i)} > 0.5?$$

"α_i" are fixed

Data

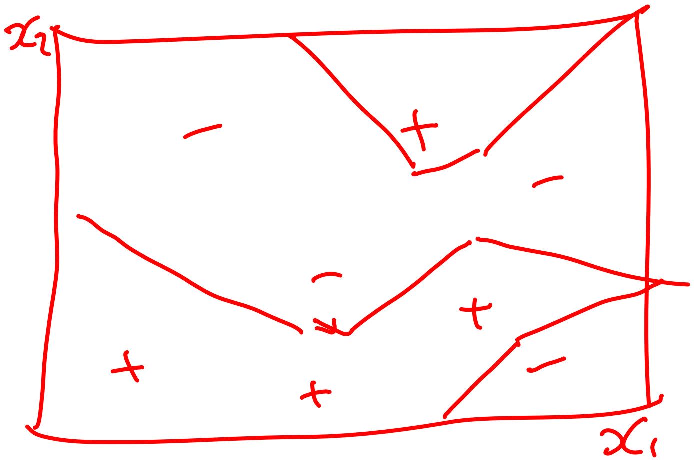
<$x_1,\ldots,x_n,y$>

Classify as

$$sign\left(\sum_i y_i K(\mathbf{x},\mathbf{x}_i) - 0.5\sum_i K(\mathbf{x},\mathbf{x}_i)\right)$$

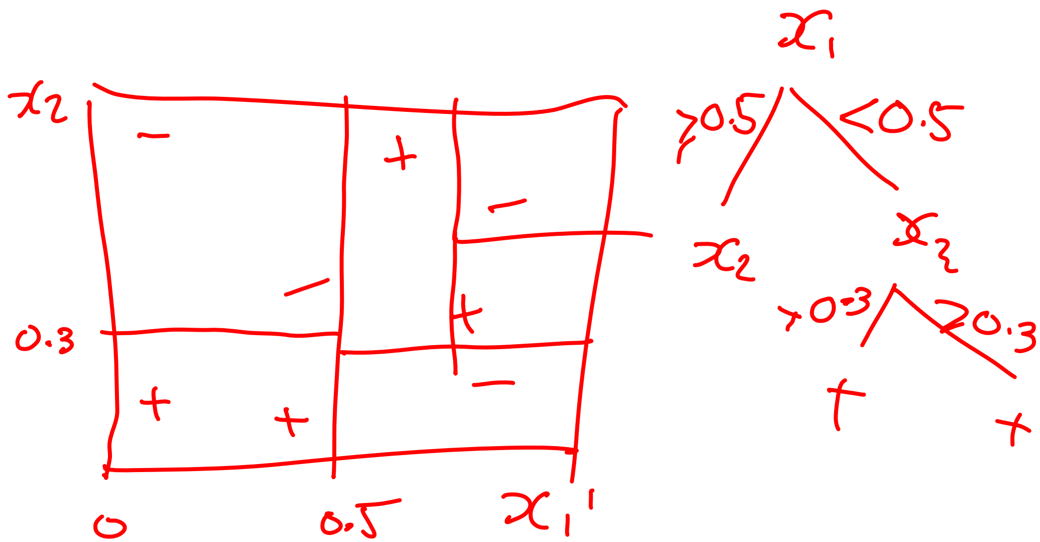
Instance-based learning versus Decision trees

1-Nearest neighbor



Voronoi split

Decision trees

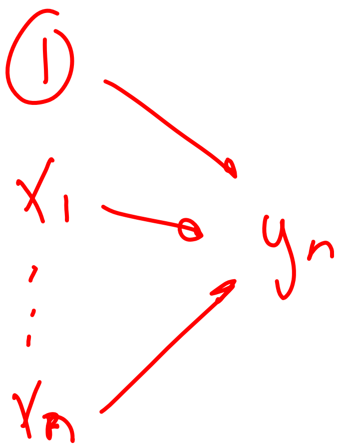


axis alligned split

Logistic regression versus Neural nets

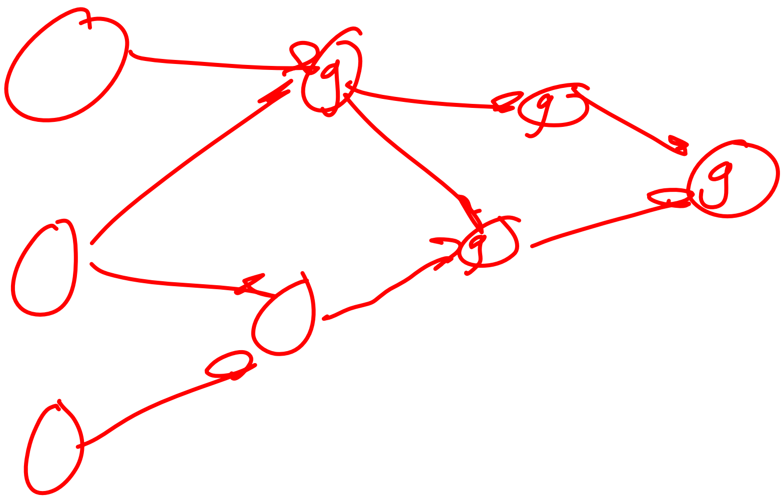
$$g(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{-(w_0 + \sum_i w_i x_i)}}$$

Logistic regression



loss: log loss

Neural Nets

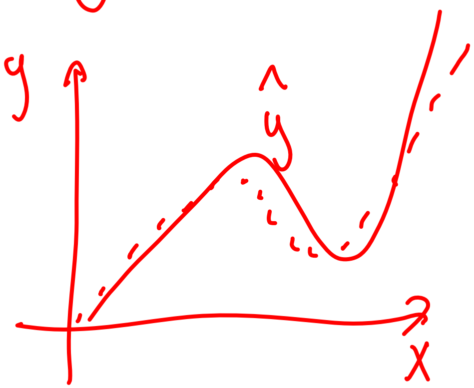


loss: sum-squared error

Linear regression versus Kernel regression

Linear Regression

$$\hat{y} = w_0 + \sum_i w_i f_i(x)$$



Kernel regression

$$\hat{y} = \frac{\sum_j w_j y_j}{\sum_j w_j}$$

$$w_j = K(x, x_j)$$

$K(\cdot, \cdot)$ has some parameters, gradient descent

Kernel-weighted linear regression

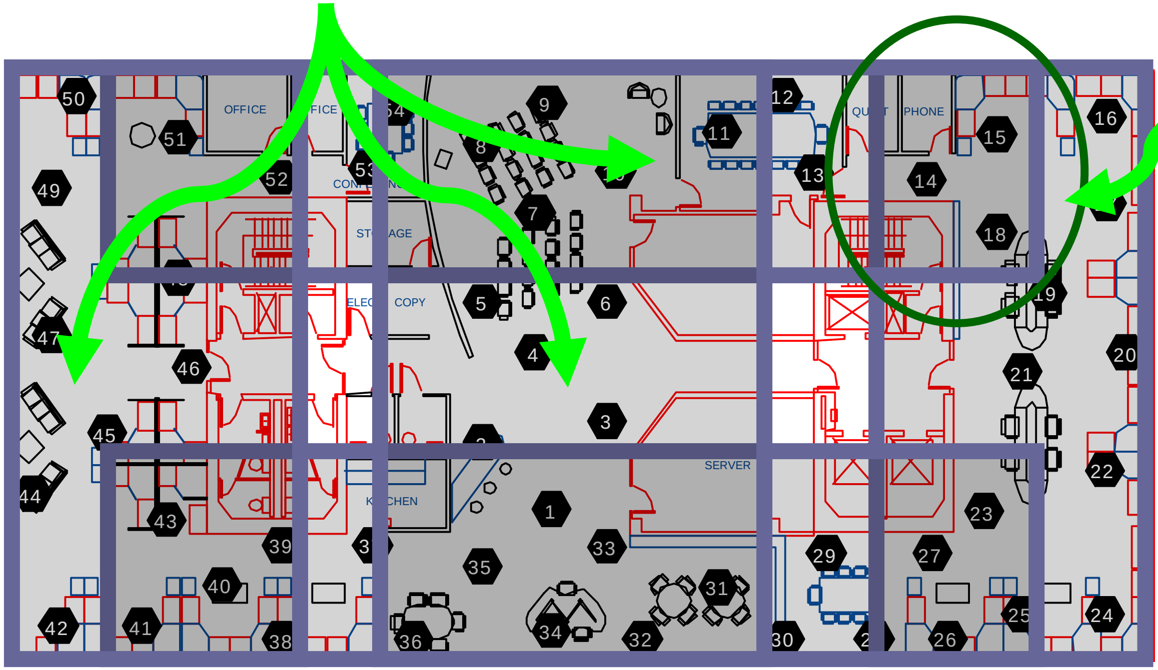
Combine:

linear regression with kernels

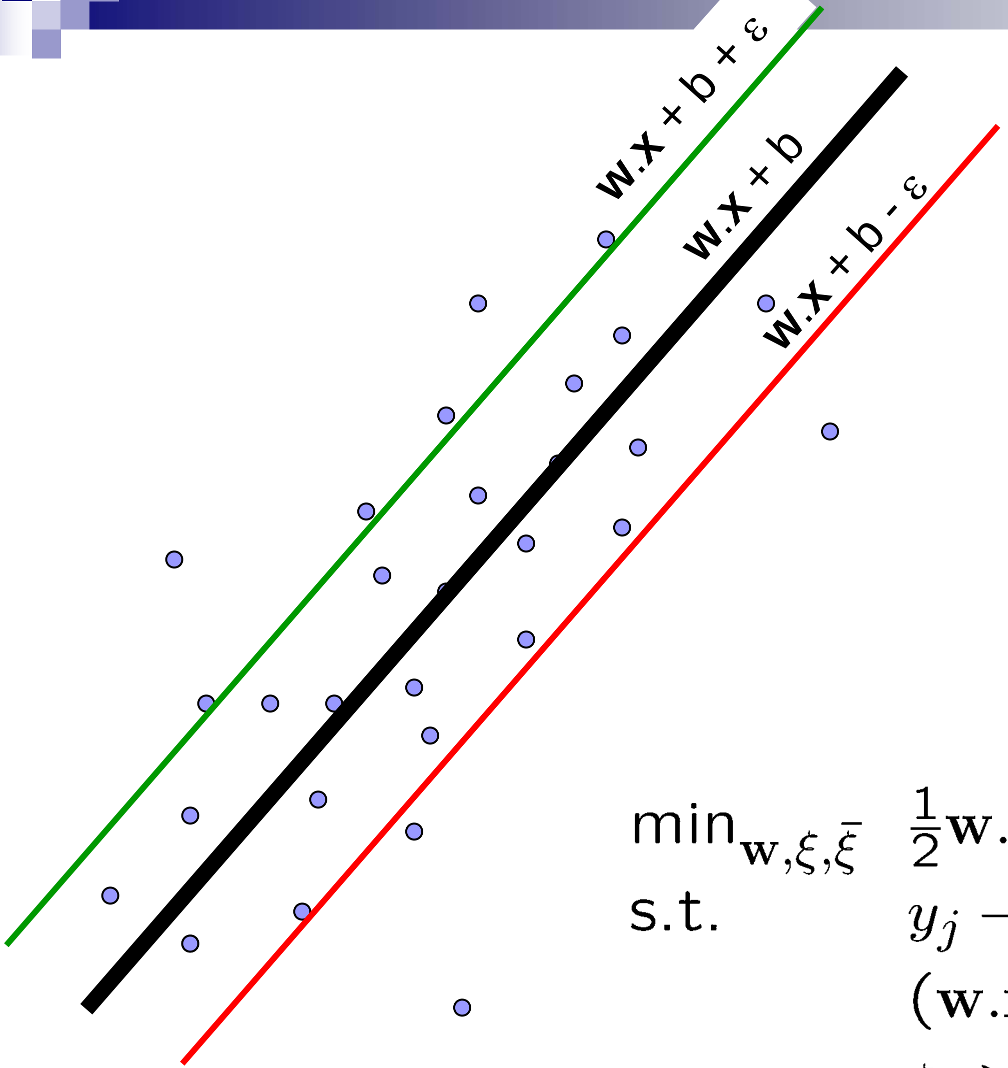
Kernel-weighted linear regression

Local basis functions for each region

Kernels average between regions



SVM regression

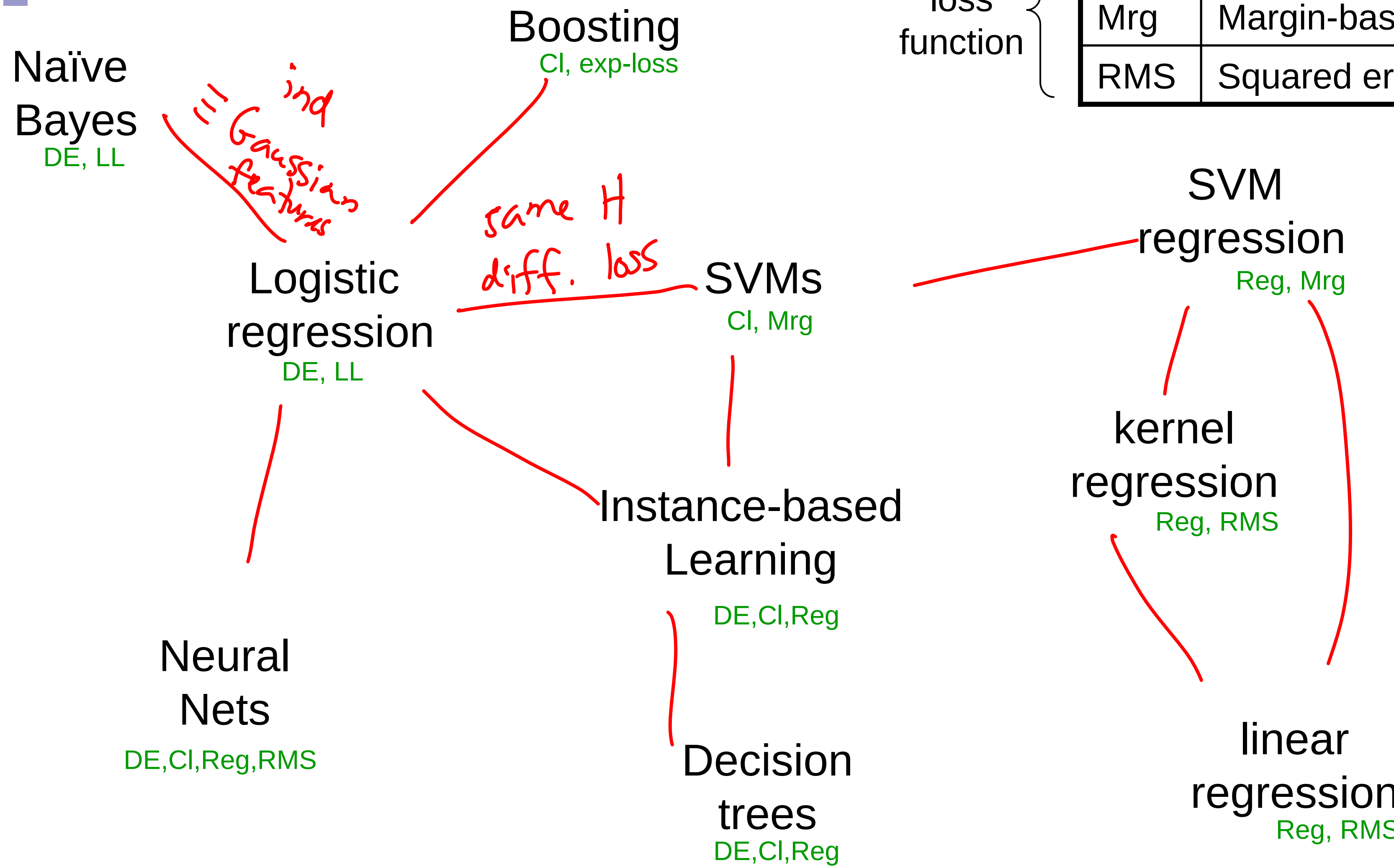


$$
\begin{array}{ll}
\min_{\mathbf{w},\xi,\bar{\xi}} & \frac{1}{2}\mathbf{w}.\mathbf{w} + C\sum_{j=1}^{m}(\xi_j + \bar{\xi}_j) \\
\text{s.t.} & y_j - (\mathbf{w}.\mathbf{x}_j + b) \leq \epsilon + \xi_j \\
 & (\mathbf{w}.\mathbf{x}_j + b) - y_j \leq \epsilon + \bar{\xi}_j \\
 & \xi_j \geq 0, \quad \bar{\xi}_j \geq 0, \quad \forall j
\end{array}
$$

BIG PICTURE

(a few points of comparison)

learning task / loss function		
learning task	DE	density estimation
learning task	Cl	Classification
learning task	Reg	Regression
loss function	LL	Log-loss/MLE
loss function	Mrg	Margin-based
loss function	RMS	Squared error



This is a very incomplete view!!!