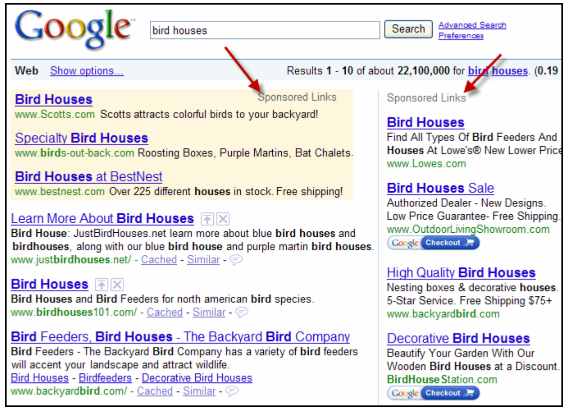
Primer on Analysis of Experimental Data and Design of Experiments

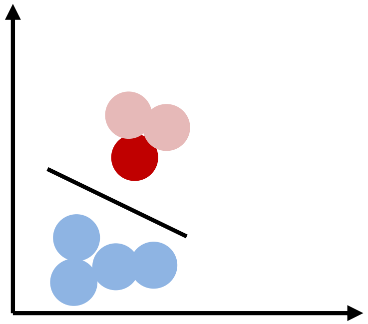
Lecture 12. Basics of Machine Learning

Muhammad A. Alam
alam@purdue.edu

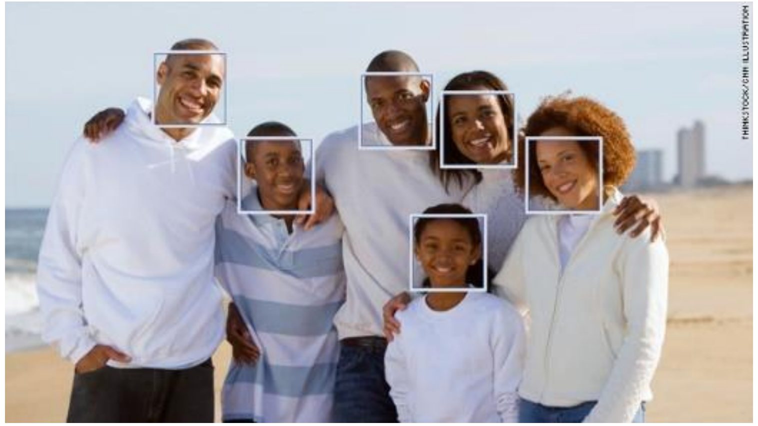
Classification problem in big data

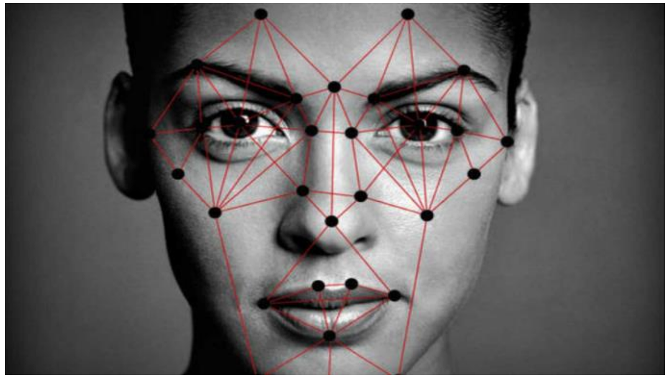
Advertisement
Recommendation





Facial Recognition
Voice Recognition
Spam Filtering





Analysis of big data

Data

Data Processing

Distance Metrics(2.1)

Sampling (2.2)

Dimensionality Reduction (2.3);

PCA (2.3.1) and

SVD (2.3.2)

Model Leanring

Supervised

Regression, Classifications

kNN (3.1.1), Decision Trees (3.1.2), Rules (3.1.3), Bayesian Classification (3.1.4), Logistic Regression (3.1.5), SVM (3.1.5), and ANN (3.1.7)

Unsupervised Learning

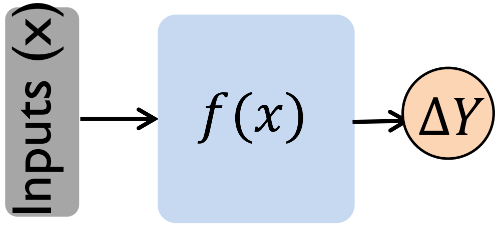
Associated Rules, Matrix Completion, Clustering)

k-means (4.1.1), Density-based (4.1.2), Message passing(4.1.2), Hierarchical (4.1.2), LDA (4.1.2), Bayesian Non-parametric (4.1.2), LSH (4.1.2)

Testing and Validation

Evaluating classifiers (3.3)

Machine Learning Introduced



$$y = f(x)$$

y: Pass, fail
y: A, B, C, D, E
Y: grade points.

$f(x)$ … Physics
$f(x)$ … Statistical curve fitting
$f_{\max}(x)$ …. Design of expt

From the headlines

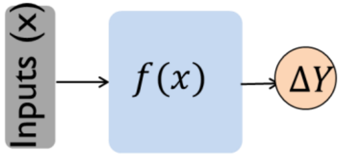
- Microsoft AI beats humans at speech recognition (TechNewsWorld)
- More accurate, fluent sentences in google translate (Barak Turovsky, lead Google Translate)
- AlphaGo: gaming that beats human (deepmind.com)
- Self driving cars (google, ….)
- Image recognition and so on …

Outline

1. Machine learning is an algorithm for “fast” curve fitting
2. Machine learning and classification: Example 1
3. Machine learning and classification: Example 2
4. Any function can be represented by machine learning approach
5. Conclusions

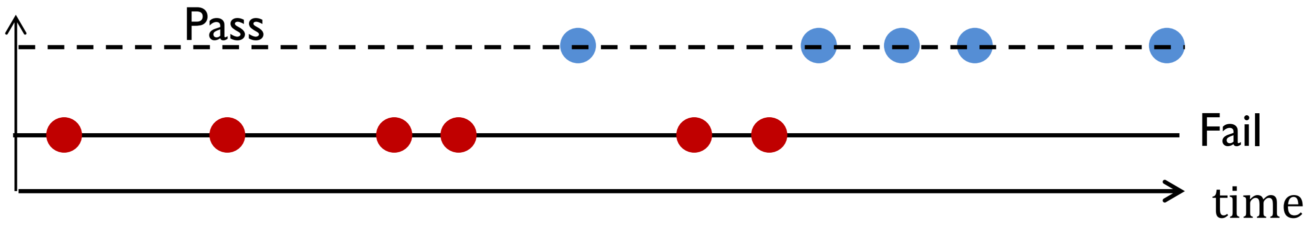
A 1D classification problem

Input: How many hours studied;
output: if they passed or failed
Goal: A "machine learning" function f(.)

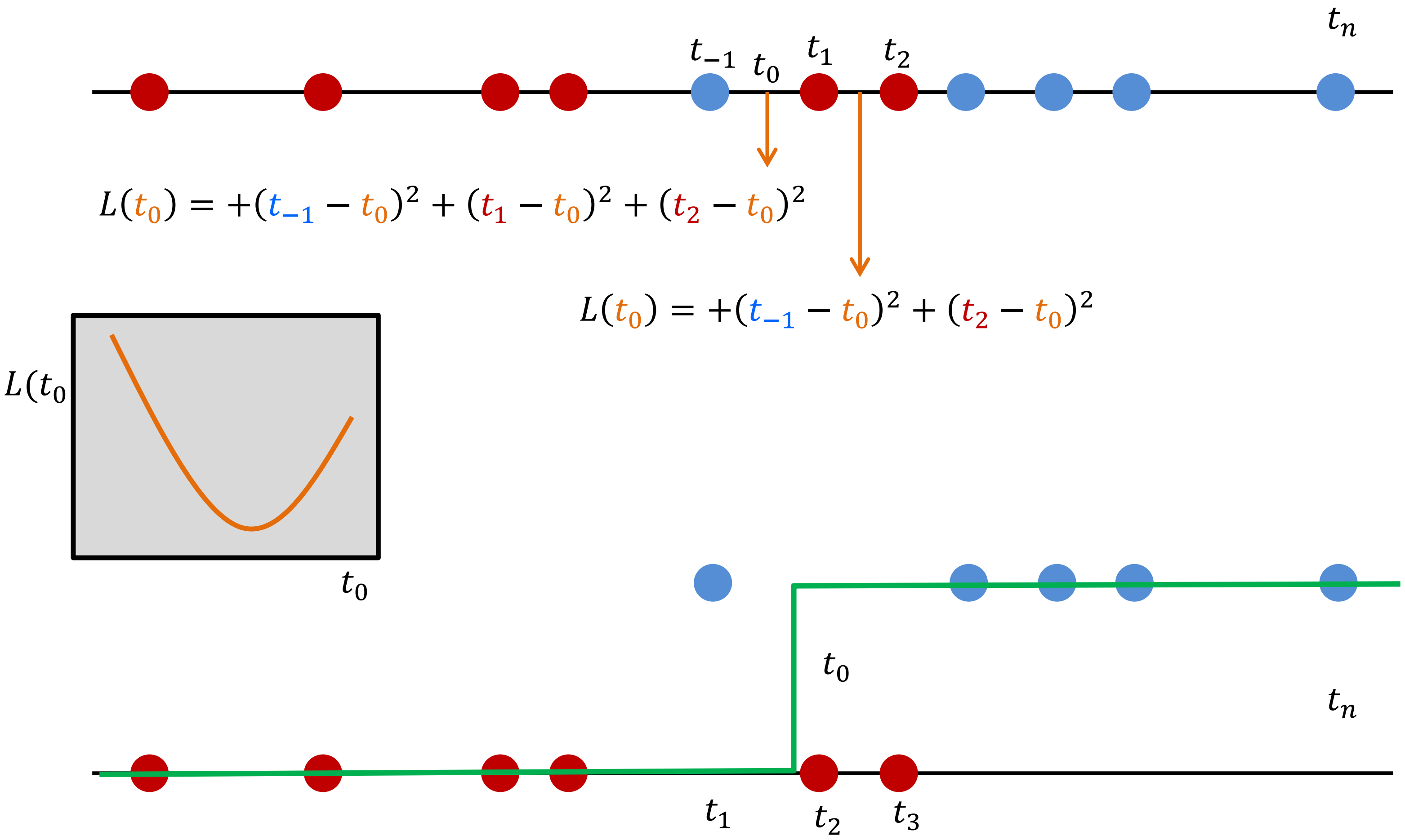


Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1



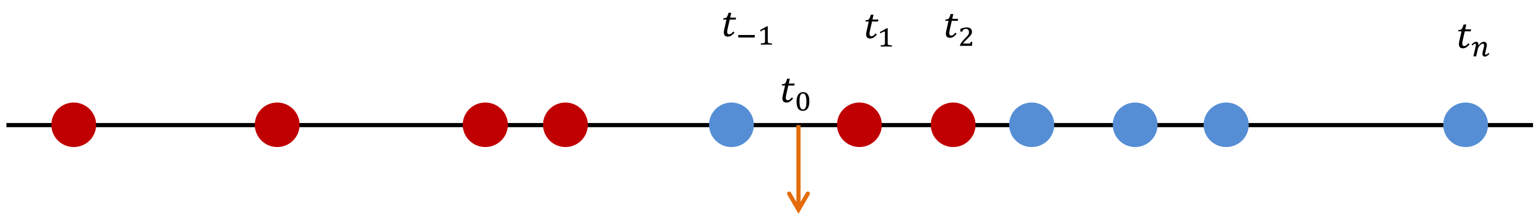
Classification: Loss function



$$L(t_0) = +(t_{-1} - t_0)^2 + (t_1 - t_0)^2 + (t_2 - t_0)^2$$

$$L(t_0) = +(t_{-1} - t_0)^2 + (t_2 - t_0)^2$$

Aside: Loss function vs. curve fitting



$L(t_0) = +(t_{-1} - t_0)^2 + (t_1 - t_0)^2 + (t_2 - t_0)^2$

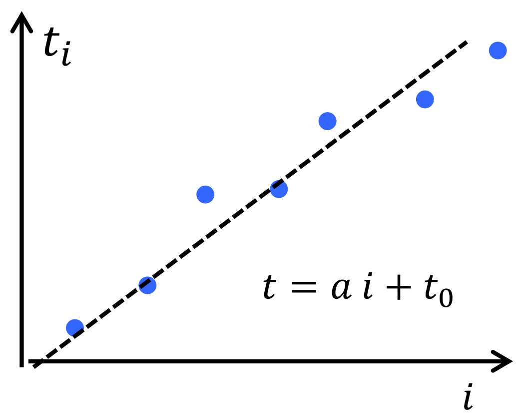
$c_i = 1$ if misclassified

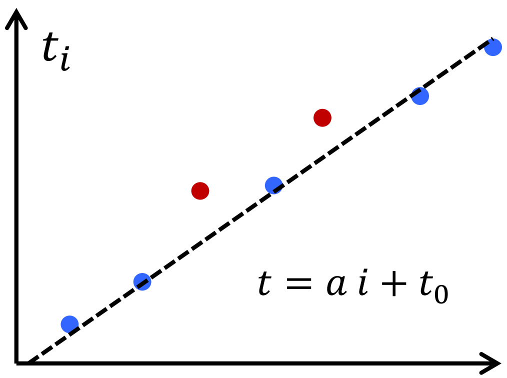
c_i defined by data quality

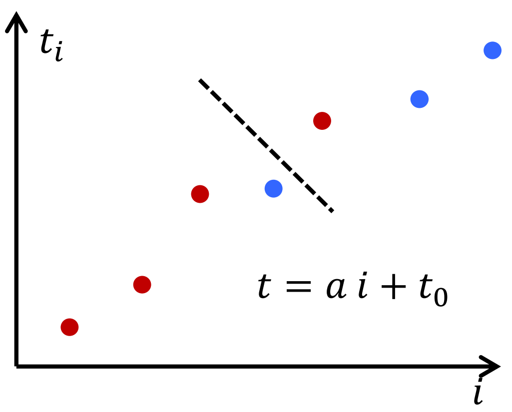
$E = \Sigma_{i=1\ldots N}\,(t_i - t)^2$

$E = \Sigma_{i=1\ldots N}\,c_i\,(t_i - t)^2$

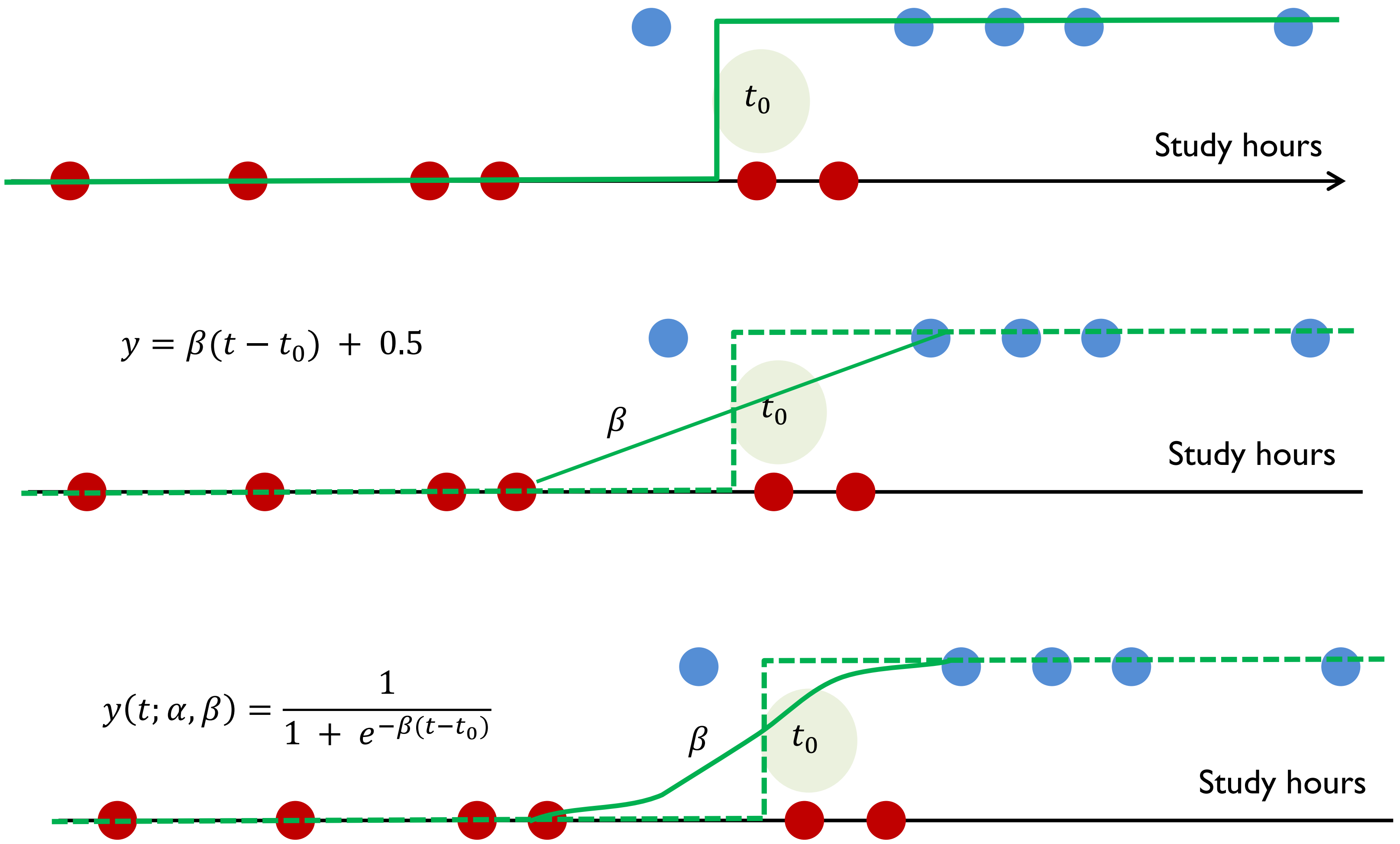
$L = \Sigma_{i=1\ldots N}\,c_i\,(t_i - t)^2$







Classification: fitting the function



Classification by sigmoidal function
A Wikipedia Example

$$\sigma(t;\alpha,\beta) == \frac{1}{1\ +\ e^{-\beta(t-t_0)}} = \frac{1}{1\ +\ e^{-(\alpha t\ +\beta\ )}}$$

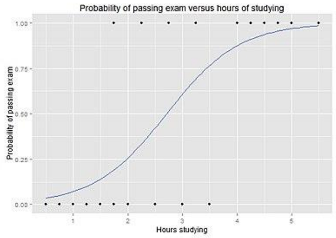
$$\sigma(t;\alpha,\beta) == \frac{1}{1\ +\ e^{-1.505(t-2.71)}}$$

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

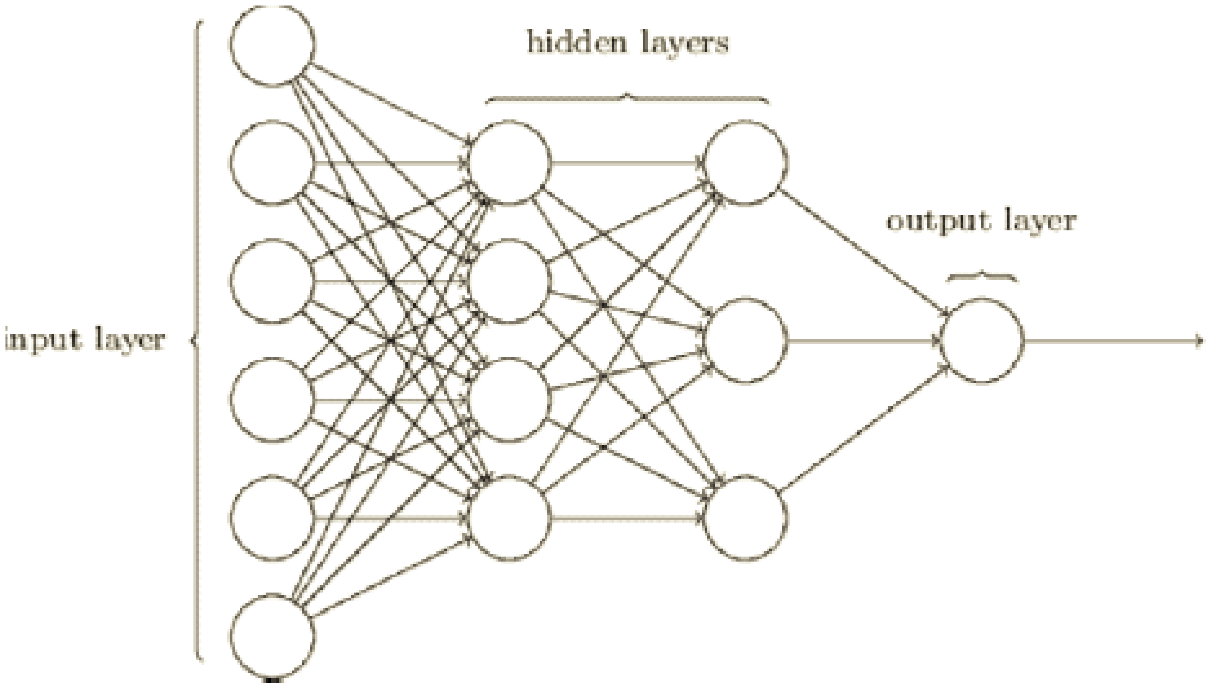
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

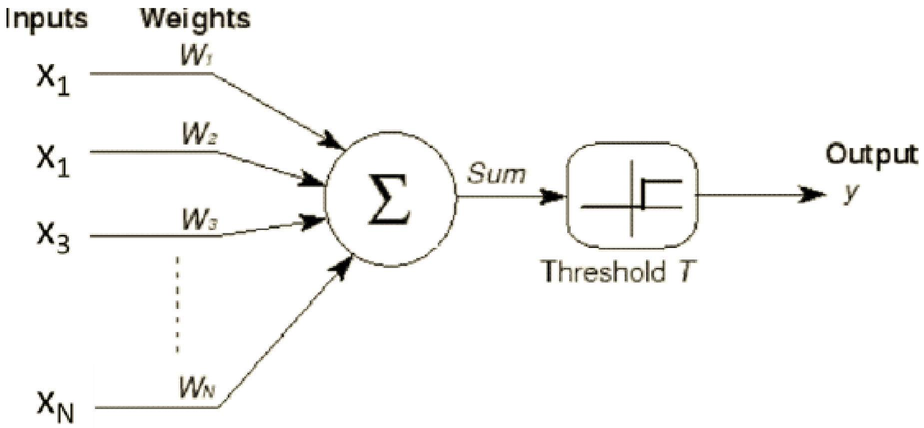
$$\ln(\sigma^{-1} - 1) = -1.505\ \mathrm{t} -\ 4.078 = -1.505\ (t - 2.71)$$

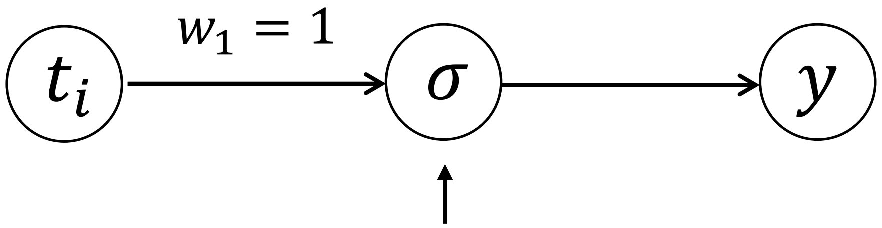
	Coefficient	Std. Error	z-value	P-value
Intercept	−4.0777	1.7610	−2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167



Our first machine learning circuit







$$\sigma(t;\alpha,\beta) == \frac{1}{1 + e^{-1.505(t-2.71)}}$$

Deriving the Loss function Coefficients

$$L_0(\alpha,\beta) \; = \prod_{i=1\ldots N} \sigma_i(\alpha,\beta)^{y_i} \times (1-\sigma_i(\alpha,\beta))^{1-y_i}$$

Compare with MLE where $L_0 = \prod_{i=1\ldots N} \; f_i$

Appropriate for binary classification

$$-L(\alpha,\beta) \; = \sum_{i=1\ldots N} [y_i \ln\big(\sigma_i(\alpha,\beta)\big) + (1-y_i)\ln(1-\sigma_i(\alpha,\beta)]$$

$dL/d\alpha \; = 0$ and $dL/d\beta \; = \; 0$ determines α and β

One input: Numerical Example

α, β

alpha	1.5046	beta	4.077
y	x	s	L
0	0.5	0.034733	0.965267
0	0.75	0.049805	0.950195
0	1	0.070936	0.929064
0	1.25	0.100088	0.899912
0	1.5	0.139422	0.860578
0	1.75	0.190934	0.809066
1	1.75	0.190934	0.190934
0	2	0.255822	0.744178
1	2.25	0.333666	0.333666
0	2.5	0.421773	0.578227
1	2.75	0.515158	0.515158
0	3	0.607496	0.392504
1	3.25	0.692738	0.692738
0	3.5	0.76658	0.23342
1	4	0.874506	0.874506
1	4.25	0.91032	0.91032
1	4.5	0.936654	0.936654
1	4.75	0.955632	0.955632
1	5	0.969112	0.969112
1	5.5	0.985201	0.985201
		sum (L)	14.72633

Binary Outcome

$$\sigma(\alpha,\beta) = \left(1+\exp\left(-(\alpha x-\beta)\right)\right)^{-1}$$

C5=1/(1+EXP(-(B1*B5- D1)

$$L_i = \left[y_i \ln\left(\sigma_i(\alpha,\beta)\right) + (1-y_i)\ln(1-\sigma_i(\alpha,\beta)\right]$$

D5=(A5*C5) +((1-A5)*(1-C5))

$$-L(\alpha,\beta) = \sum_{i=1\ldots N} L_i(\alpha,\beta)$$

D5=SUM (D5:D24)

Homework: One input optimization

1. Excel-based HW for understanding the fitting process.

2. Use logistic calculator to optimize the coefficient
http://statpages.info/logistic.html http://statpages.info/logistix.html

Data

0.5,0
0.75,0
1.0,0
1.25,0
1.5,0
1.75,0
1.75,1
2.0,0
2.25,1
2.5,0
2.75,1
3.0,0
3.25,1
3.5,0
4.0,1
4.25,1
4.5,1
4.75,1
5.0,1
5.5,1

Descriptives...

10 cases have Y=0; 10 cases have Y=1.

Variable	Avg	SD
1	2.7875	1.4690

Iteration History...
-2 Log Likelihood = 27.7259 (Null Model)
-2 Log Likelihood = 25.9205
-2 Log Likelihood = 23.1187
-2 Log Likelihood = 20.3710
-2 Log Likelihood = 18.2717
-2 Log Likelihood = 16.9599
-2 Log Likelihood = 16.3181
-2 Log Likelihood = 16.1022
-2 Log Likelihood = 16.0626
-2 Log Likelihood = 16.0598
-2 Log Likelihood = 16.0598
-2 Log Likelihood = 16.0598 (Converged)

Overall Model Fit...
Chi Square= 11.6661; df=1; p= 0.0006
Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

Variable	Coeff.	StdErr	p	O.R.	Low	High
1	1.5046	0.6287	0.0167	4.5026	1.3131	15.4393
Intercept	-4.0777	1.7610	0.0206			

Descriptives...

Predicted Probability of Outcome, with 95% Confidence Limits...

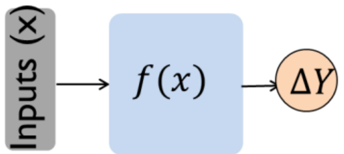
X	Y	Prob	Low	High
0.5000	0	0.0347	0.0020	0.3914
0.7500	0	0.0498	0.0038	0.4157
1.0000	0	0.0709	0.0073	0.4424
1.2500	0	0.1000	0.0136	0.4722
1.5000	0	0.1393	0.0249	0.5063
1.7500	0	0.1908	0.0442	0.5460
1.7500	1	0.1908	0.0442	0.5460
2.0000	0	0.2557	0.0749	0.5933
2.2500	1	0.3335	0.1189	0.6498
2.5000	0	0.4216	0.1745	0.7155
2.7500	1	0.5150	0.2349	0.7860
3.0000	0	0.6074	0.2930	0.8524
3.2500	1	0.6926	0.3444	0.9062
3.5000	0	0.7665	0.3885	0.9443
4.0000	1	0.8744	0.4599	0.9827
4.2500	1	0.9103	0.4897	0.9908
4.5000	1	0.9366	0.5168	0.9951
4.7500	1	0.9556	0.5418	0.9975
5.0000	1	0.9691	0.5651	0.9987
5.5000	1	0.9852	0.6079	0.9997

Outline

1. Machine learning is an algorithm for “fast” curve fitting
2. Machine learning and classification: Example 1
3. Machine learning and classification: Example 2
4. Any function can be represented by machine learning approach
5. Conclusions

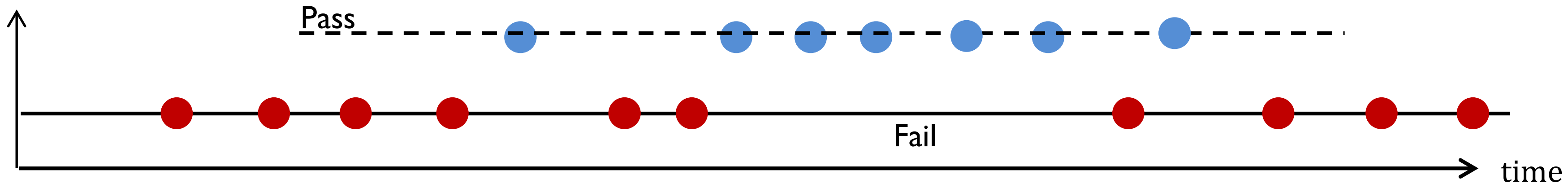
Generalized 1D classification problem

Input: How many hours studied;
output: if they passed or failed
Goal: A "machine learning" function f(.)

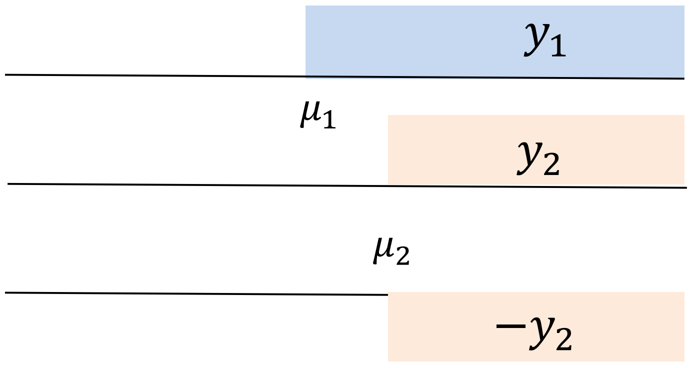


Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0

Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50	5.75	6.0	6.25	6.5	6.75	7.0
Pass	1	0	1	0	1	1	1	1	1	1	0	1	0	0	0	0



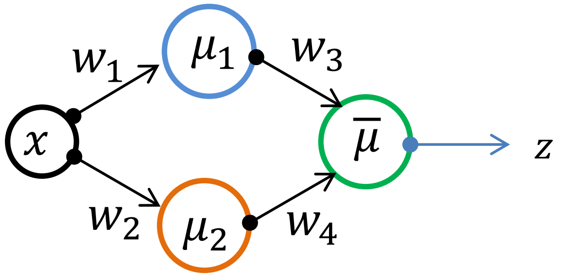
A bit more complex classification



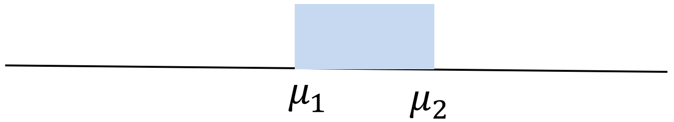
$$y_1 \;=\; 1\,/\,(1 + \;\exp(-(w_1 x - \mu_1)/\sigma)$$

$$y_2 \;=\; 1/(1 + \exp(-\,(w_2 x - \mu_2)/\sigma)$$

$$w_1 = 1, \quad w_2{=}1$$



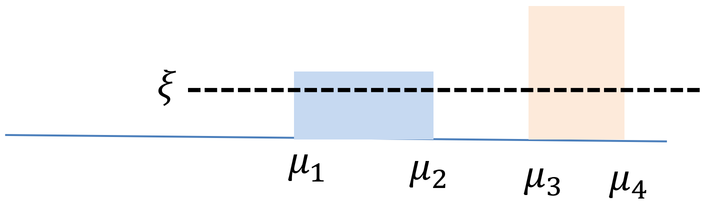
$$\overline{\mu} = \;(\mu_1 \; + \mu_2)/2$$

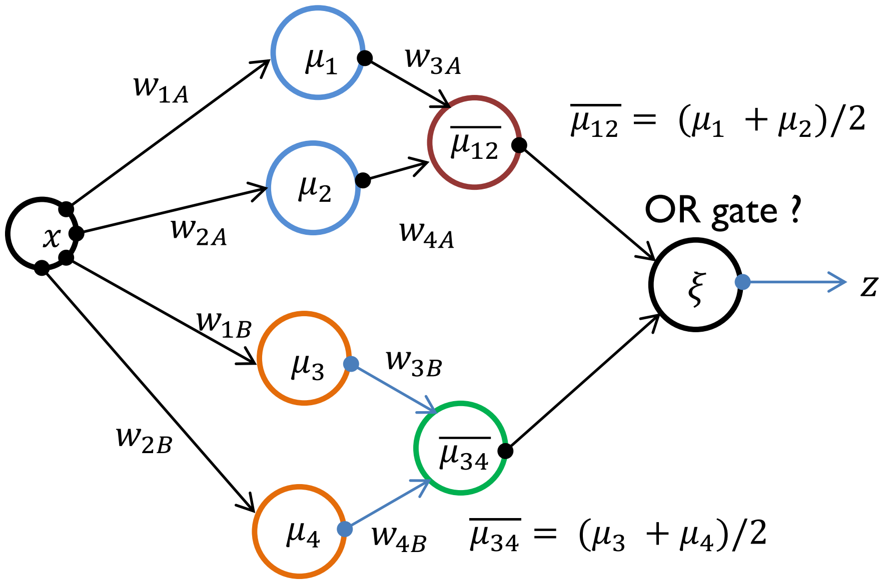


$$z = \;1\,/\,(1 + \;\exp(-(w_3 y_1 + w_4 y_2 - \mu_a)/\sigma)$$

$$w_3 = 1, w_4 = -1$$

Any f(x) can be represented by a ML network





$$y_{1A} = 1 / (1 + \exp(-(w_{1A}x - \mu_{1A})/\sigma)$$

$$y_{2A} = 1/(1 + \exp(-(w_{2A}x - \mu_{2A})/\sigma)$$

$$z_A = 1 / (1 + \exp(-(w_{3A}y_1 + w_{4A}y_2 - \overline{\mu_{12}})/\sigma)$$

$$y_{1B} = 1 / (1 + \exp(-(w_{1B}x - \mu_{1B})/\sigma)$$

$$y_{2B} = 1/(1 + \exp(-(w_{2B}x - \mu_{2B})/\sigma)$$

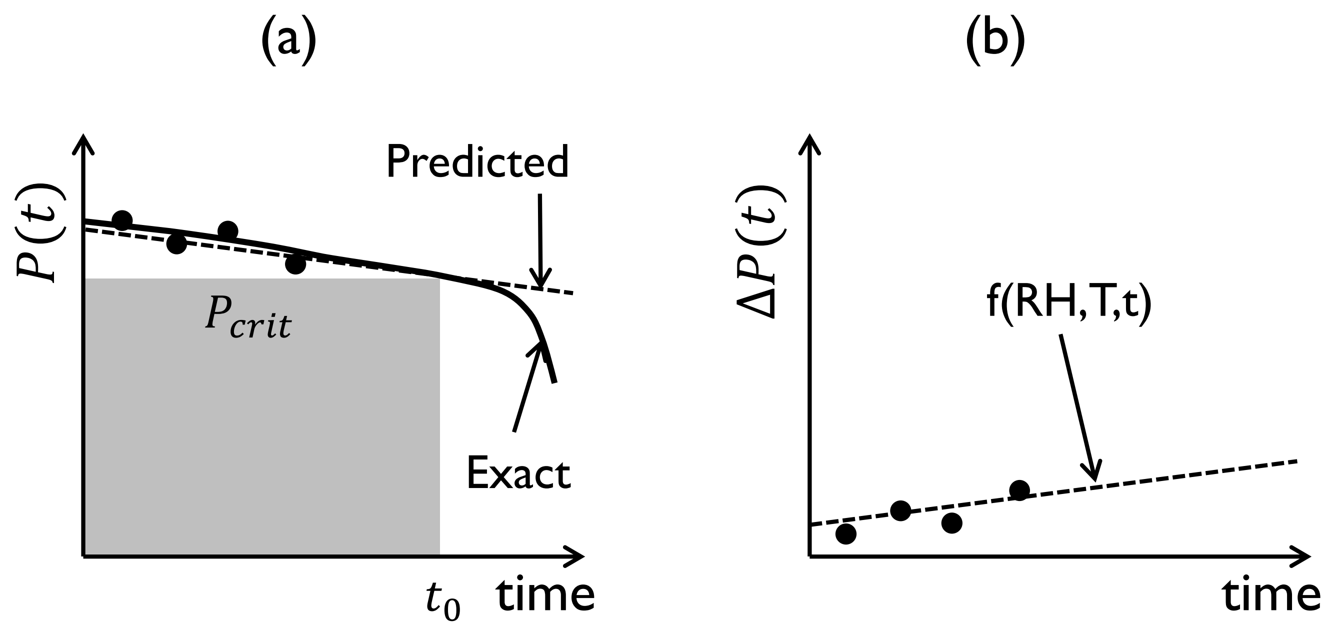
$$z_B = 1 / (1 + \exp(-(w_3y_1 + w_4y_2 - \overline{\mu_{12}})/\sigma)$$

$$z = 1 / (1 + \exp(-(z_A + z_B - \xi)/\sigma)$$

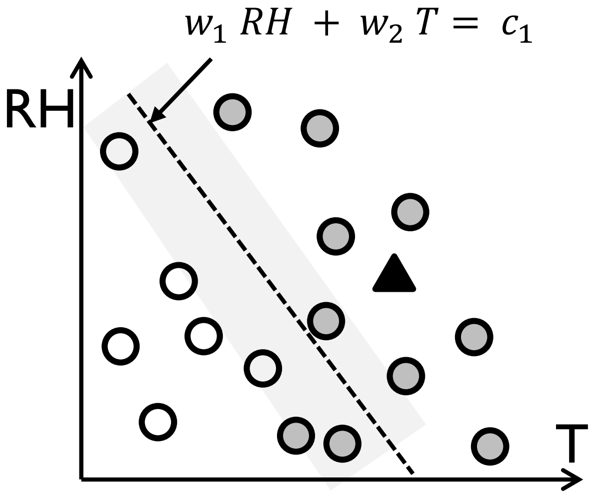
$$w_{1A} = w_{2A} = w_{1B} = w_{\{2B)} = 1, w_2{=}1$$

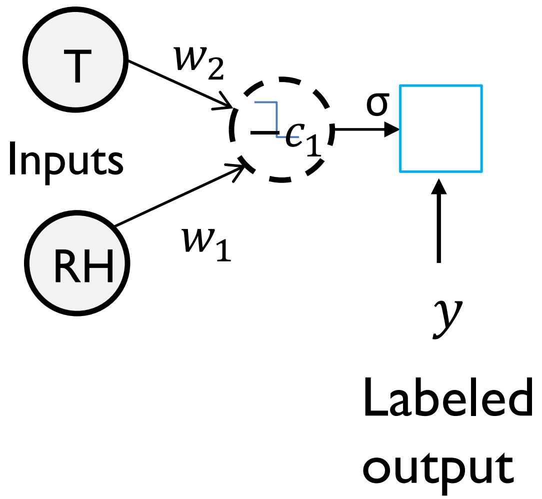
$$w_{3A} = w_{3B} = 1, \quad w_{4A} = w_{4B} = -1$$

Reliability of Solar Farms …



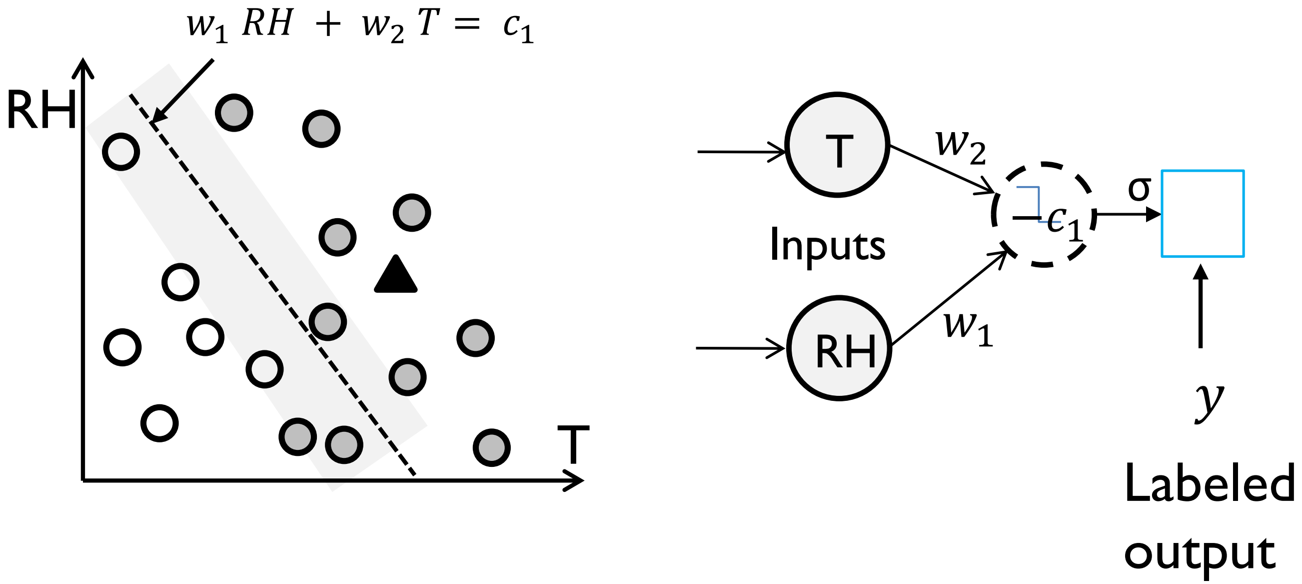
…. represented by two input ANN





$$\sigma(w_1, w_2, c) = \frac{1}{1 + \exp(-(w_1 T + w_2\,RH - c_1)/\sigma)}$$

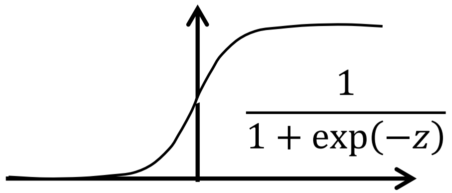
Training by backpropagation



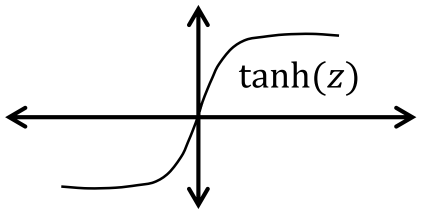
Algorithms by computer scientists
We only have straight lines, hence many layers

Aside: Transition Functions

sigmoid

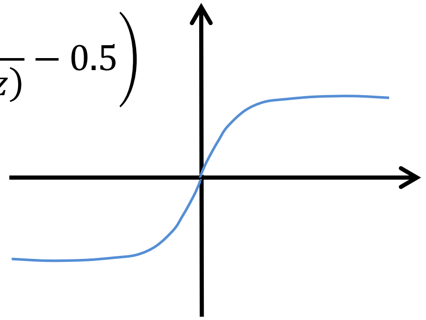


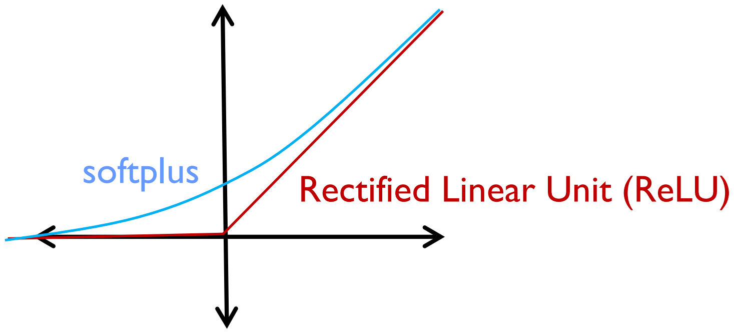
tanh



Shifted sigmoid

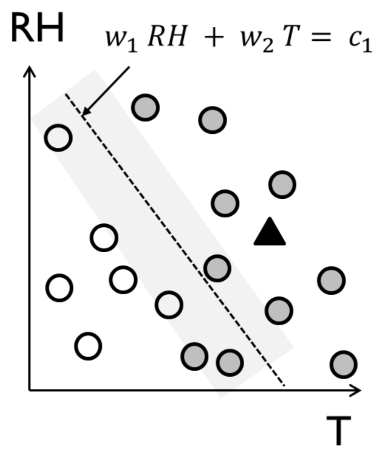
$$2\left(\frac{1}{1+\exp(-z)} - 0.5\right)$$



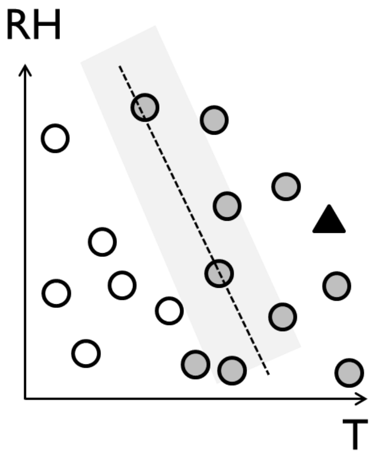


Sigmoid/tanh emphasizes points close to transition

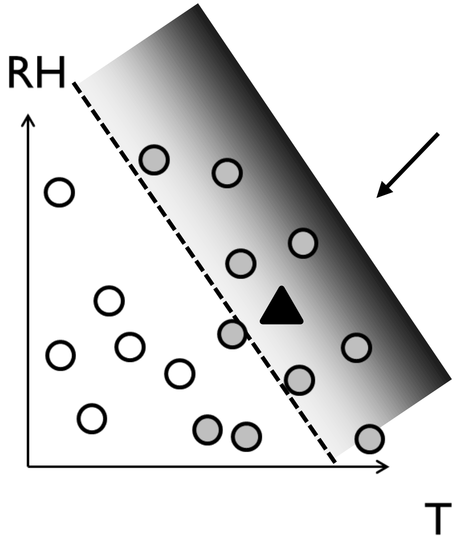
Aside: Different transition functions



Sigmoidal



Support Vector Machine



LiRu

Conclusions

1. Classification of data is an important statistical problem with applications in advertisement, recommendation, etc.

2. Machine learning is an empirical (and easily generalizable) multi-parameter curve fitting process. While SVD is more powerful, machine learning applies to larger datasets.

3. Any function can be represented by a machine learning algorithm. The definition of loss function and quick calculation of coefficients are the key issues.

4. We have focused on one or two input systems. The problem is easily generalized.

Review Questions

1. What is the difference between a sigmoid function and a tanh function?
2. Why can a XOR not be implemented by a single neuron or perceptron?
3. If the weights of the input to a OR-neuron is 0.6 and 1.2, what should be its threshold?
4. What does the support vectors of a support vector machine (SVM) refer to?
5. In what ways is a SVM is better than a sigmoidal transition function?
6. How does a random forest model compare with that of neural network model?
7. What is a loss function?
8. How does the sigmoid or tanh transformation of the original data reduces the sensitivity of accidental misclassification of the data?

References

1 layers all continuous functions
2 layers all functions even with discontinuity

Kolmogorov, Andrei Nikolaevich. "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition." Doklady Akademii Nauk. Vol. 114. No. 5. Russian Academy of Sciences, 1957.

Cybenko, George. "Approximation by superpositions of a sigmoidal function." Mathematics of control, signals and systems 2.4 (1989): 303-314