Artificial Intelligence
EDA132
Lecture 6: Machine Learning

Pierre Nugues

Lund University
Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

February 3, 2017

Linear Classifiers

Similar to decision trees, discriminant analysis provides another set of techniques to learn classifiers from data sets.
This time, the objects will be represented by a vector of numerical parameters, often called the **features**.
Linear classification methods operate in an n-dimensional space with a vector space equal to the number of features.
Linear classifiers include:

1. Perceptron
2. Logistic regression
3. Support vector machines

Problem Formulation

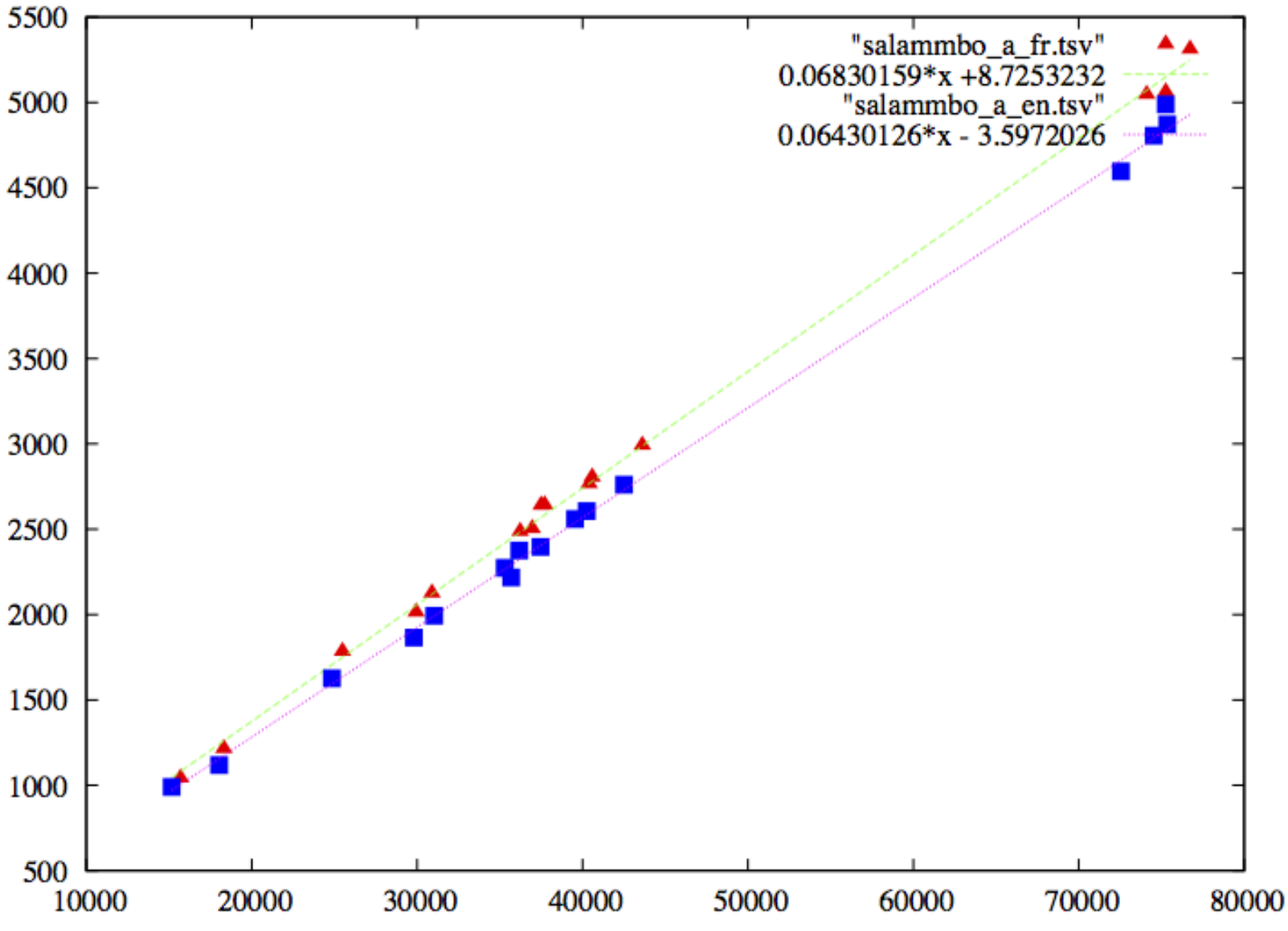
We can formulate the problem as finding lines automatically to:

1. Fit a data set (regression). The output is a real continuous variable.
2. Separate classes inside a data set (discriminant analysis, classification). The output is a discrete set of symbols, the classes, for instance {0, 1} or {negative, positive}.
   Here, we will consider the binary case only with two classes. Multiclass classification is a generalization of it.

Linear Regression: Letter Frequencies in *Salammbô*

Chapter	French		English	
	# characters	# *A*	# characters	# *A*
Chapter 1	36,961	2,503	35,680	2,217
Chapter 2	43,621	2,992	42,514	2,761
Chapter 3	15,694	1,042	15,162	990
Chapter 4	36,231	2,487	35,298	2,274
Chapter 5	29,945	2,014	29,800	1,865
Chapter 6	40,588	2,805	40,255	2,606
Chapter 7	75,255	5,062	74,532	4,805
Chapter 8	37,709	2,643	37,464	2,396
Chapter 9	30,899	2,126	31,030	1,993
Chapter 10	25,486	1,784	24,843	1,627
Chapter 11	37,497	2,641	36,172	2,375
Chapter 12	40,398	2,766	39,552	2,560
Chapter 13	74,105	5,047	72,545	4,597
Chapter 14	76,725	5,312	75,352	4,871
Chapter 15	18,317	1,215	18,031	1,119
Total	619,431	42,439	608,230	39,056

Fitting the Data



Fitting Equations

In a two-dimensional space, the straight line fitting the points is given by:

$$\hat{y} = mx + b,$$

Minimize the loss between:

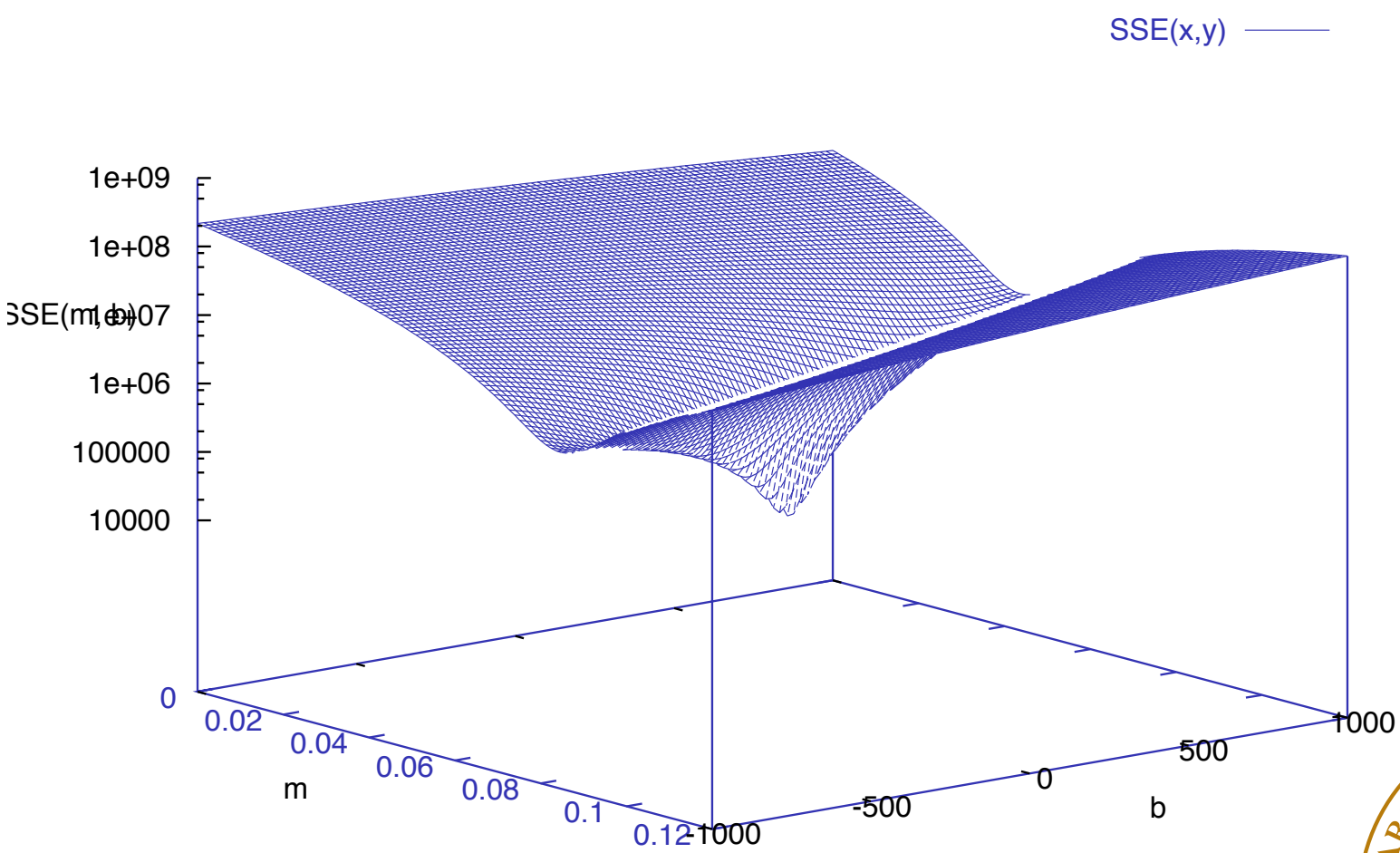
- The set of q observations: $\{(x_i, y_i)\}_{i=1}^{q}$ and
- A perfect linear alignment: $\{(x_i, f(x_i))\}_{i=1}^{q}$, where $f(x_i) = mx_i + b$.

The loss is traditionally modeled by the quadratic error (Legendre 1805):
$L_2 = \sum_{\text{Set of points}} (y - \hat{y})^2$.
For q points (x_i, y_i):

$$Loss(b, m) = \sum_{i=1}^{q} (y_i - (mx_i + b))^2$$

The objective of regression is to find the weights that minimize the loss.

Visualizing the Loss



Minimizing the Loss

The loss function is convex and has a unique minimum.
The loss reaches a minimum when the partial derivatives are zero:

$$\frac{\partial Loss}{\partial m} = \sum_{i=1}^{q} \frac{\partial}{\partial m}(y_i - (mx_i + b))^2 = -2\sum_{i=1}^{q} x_i(y_i - (mx_i + b)) = 0$$

$$\frac{\partial Loss}{\partial b} = \sum_{i=1}^{q} \frac{\partial}{\partial b}(y_i - (mx_i + b))^2 = -2\sum_{i=1}^{q} (y_i - (mx_i + b)) = 0$$

Fitting Equations (II)

This yields:

$$\begin{aligned} b &= \bar{y} - m\bar{x} \\ \sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{q} x_i^2 - qb\bar{x} &= \sum_{i=1}^{q} x_i y_i - m\sum_{i=1}^{q} x_i^2 - q\bar{x}(\bar{y} - m\bar{x}) = 0 \end{aligned}$$

$$m = \frac{\sum_{i=1}^{q} x_i y_i - q\bar{x}\bar{y}}{\sum_{i=1}^{q} x_i^2 - q\bar{x}^2} \quad \text{and} \quad b = \bar{y} - m\bar{x},$$

with

$$\bar{x} = \frac{1}{q}\sum_{i=1}^{q} x_i \quad \text{and} \quad \bar{y} = \frac{1}{q}\sum_{i=1}^{q} y_i.$$

$$\begin{aligned} \text{French:} \quad y &= 0.0683x + 8.7253 \\ \text{English:} \quad y &= 0.0643x - 3.5972 \end{aligned}$$

Regression: Notation in an N-dimensional Space

1. The feature vector (or predictors): $(1, x_1, x_2, ..., x_{n-1})$ representing the input. In our example, this corresponds to the letter frequencies: (1, 36961) in Chapter 1. The first parameter is set to 1;
2. The output (or response or target): y. In our example, the frequencies of a: 2503 in Chapter 1; the predicted output: $\hat{y} = 0.0683 \times 36961 + 8.7253 = 2533.22$
3. The squared error for one point: $(2503 - 2533.22)^2 = 30.22^2 = 913.26$
4. To denote the whole data set, we use a matrix, $\mathbf{X}$, where each row corresponds to an observation and a vector $\mathbf{y}$ for the outputs

Regression: Notation in an N-dimensional Space (II)

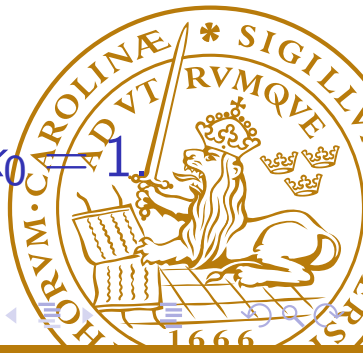
The least-squares method minimizes the sum of the squared errors for all the observations through optimal values of m and b. In an n-dimensional space, we have:

1. The weight vector: $(w_0, w_1, w_2, ..., w_{n-1})$, $(8.7253, 0.0683)$ for French and $(-3.5972, 0.0643)$ for English;
2. The intercept: w_0.

The hyperplane equation corresponds to the dot product of the weights by the features defined as:

$$y = \mathbf{w} \cdot \mathbf{x} = \sum_{i=0}^{n-1} w_i x_i,$$

where $\mathbf{w} = (w_0, w_1, w_2, ..., w_{n-1})$, $\mathbf{x} = (x_0, x_1, x_2, ..., x_{n-1})$, and $x_0 = 1$.



Another Technique: The Gradient Descent

The gradient descent is a numerical method to find the minimum of $f(x_0, x_1, x_2, ..., x_n) = y$, when there is no analytical solution.
Let us denote $\mathbf{x} = (x_0, x_1, x_2, ..., x_n)$
We derive successive approximations to find the minimum of f:

$$f(\mathbf{x}_1) > f(\mathbf{x}_2) > ... > f(\mathbf{x}_k) > f(\mathbf{x}_{k+1}) > .. > min$$

Points in the neighborhood of $\mathbf{x}$ are defined by $\mathbf{x} + \mathbf{v}$ with $||\mathbf{v}||$ small
Given $\mathbf{x}$, find $\mathbf{v}$ subject to $f(\mathbf{x}) > f(\mathbf{x} + \mathbf{v})$

The Gradient Descent (Cauchy, 1847)

Using a Taylor expansion: $f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \mathbf{v}\cdot\nabla f(\mathbf{x}) + \ldots$
The gradient is a direction vector corresponding to the steepest slope:

$$\nabla f(x_0, x_1, x_2, ..., x_n) = (\frac{\partial f}{\partial x_0}, \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}).$$

Steepest ascent: $\mathbf{v} = \alpha\nabla f(\mathbf{x})$, steepest descent: $\mathbf{v} = -\alpha\nabla f(\mathbf{x})$, where $\alpha > 0$.
We have then: $f(\mathbf{x} - \alpha\nabla f(\mathbf{x})) \approx f(\mathbf{x}) - \alpha||\nabla f(\mathbf{x})||^2$.
The inequality:

$$f(\mathbf{x}) > f(\mathbf{x} - \alpha\nabla f(\mathbf{x}))$$

enables us to move one step down to the minimum.
We then use the iteration:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k).$$

Algorithm

For a data set DS, we find the minimum through a walk (iteration) down the surface.

1: **function** Regression(DS)
2: $\quad \mathbf{w} \leftarrow \mathbf{w^0}$ ▷ $\mathbf{w^0}$: Any point in the space
3: $\quad$ **while** $||\nabla Loss(\mathbf{w})|| > \varepsilon$ **do**
4: $\quad\quad \mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla Loss(\mathbf{w})$
5: $\quad$ **return w**

α is a positive number either constant or variable.

In a Two-dimensional Space

To make the generalization easier, let us rename the straight line coefficients (b, m) as (w_0, w_1)
Given a data set DS of q examples: $DS = \{(1, x_1^j, y^j) | j : 1..q\}$, where the error is defined as:

$$\begin{aligned} SSE(w_0, w_1) &= \sum_{j=1}^{q} (y^j - \hat{y}^j)^2 \\ &= \sum_{j=1}^{q} (y^j - (w_0 + w_1 x_1^j))^2. \end{aligned}$$

The gradient with q examples:

$$\begin{aligned} \frac{\partial SSE}{\partial w_0} &= -2 \sum_{j=1}^{q} (y^j - (w_0 + w_1 x_1^j)) \\ \frac{\partial SSE}{\partial w_1} &= -2 \sum_{j=1}^{q} x_1^j \times (y^j - (w_0 + w_1 x_1^j)) \end{aligned}$$

Updates

Batch gradient descent with q examples:

$$\begin{aligned} w_0 &\leftarrow w_0 + \frac{\alpha}{q}\sum_{j=1}^{q}(y^j - (w_0 + w_1 x_1^j)) \\ w_1 &\leftarrow w_1 + \frac{\alpha}{q}\sum_{j=1}^{q} x_1^j \times (y^j - (w_0 + w_1 x_1^j)) \end{aligned}$$

Stochastic gradient descent. The updates are carried out one example at a time:

$$\begin{aligned} w_0 &\leftarrow w_0 + \alpha \cdot (y^j - (w_0 + w_1 x_1^j)) \\ w_1 &\leftarrow w_1 + \alpha \cdot x_1^j \cdot (y^j - (w_0 + w_1 x_1^j)) \end{aligned}$$

The stochastic variant is also called online learning and is usually faster.

The Gradient Descent and Linear Regression

In the general case, we want to find the regression hyperplane:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n$$

given a data set of q examples: $DS = \{(1, x_1^j, x_2^j, ..., x_n^j, y^j) | j : 1..q\}$, where the loss is defined as:

$$\begin{aligned} Loss(w_0, w_1, ..., w_n) &= \sum_{j=1}^{q} (y^j - \hat{y}^j)^2 \\ &= \sum_{j=1}^{q} (y^j - (w_0 + w_1 x_1^j + w_2 x_2^j + ... + w_n x_n^j))^2 \end{aligned}$$

We introduce $x_0 = 1$ and we compute the gradient to determine the slope:

$$\frac{\partial Loss}{\partial w_i} = -2 \sum_{j=1}^{q} x_i^j \times (y^j - (w_0 x_0^j + w_1 x_1^j + w_2 x_2^j + ... + w_n x_n^j))$$

Updates

In the **batch gradient descent**, the iteration step considers all the examples, q examples here:

$$w_i \leftarrow w_i + \frac{\alpha}{q} \cdot \sum_{j=1}^{q} x_i^j \cdot (y^j - (w_0 x_0^j + w_1 x_1^j + w_2 x_2^j + \ldots + w_n x_n^j))$$

Using a matrix notation:

$$\mathbf{w} \leftarrow \mathbf{w} + \frac{\alpha}{q} \cdot (\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal \mathbf{X}.$$

In the **stochastic gradient descent**, we carry out the updates using one example at a time:

$$w_i \leftarrow w_i + \alpha \cdot x_i^j \cdot (y^j - (w_0 x_0^j + w_1 x_1^j + w_2 x_2^j + \ldots + w_n x_n^j))$$

Using a vector notation with $\mathbf{x}^j$ and y^j, we have:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \cdot (y^j - \mathbf{x}^j \cdot \mathbf{w}) \cdot \mathbf{x}^j.$$

Definitions

An epoch is the period corresponding to one iteration over the complete data set: the q examples.

α is called the learning rate or the step size. It is a positive number that is either fixed or varies (decreases) over the epochs.

Setting α can be difficult. In the assignments, you will manually try a set of values and check the convergence.
Murphy (2012) describes other methods (for logistic regression). See Sect. 8.3.2 and 8.5.2. They are outside the scope of this course.
See also the articles published by the LIBLINEAR team:
`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`.

The Analytical Solution Again

In a two-dimensional space, linear regression has an analytical solution:

$$m = \frac{\sum_{i=1}^{q} x_i y_i - q\bar{x}\bar{y}}{\sum_{i=1}^{q} x_i^2 - q\bar{x}^2} \qquad \text{and} \qquad b = \bar{y} - m\bar{x}.$$

This can be generalized for any dimension.
Let $\mathbf{X} = (\mathbf{x}^1; \mathbf{x}^2; ...; \mathbf{x}^q)$ be our observations and $\mathbf{y}$, the output. We have:

$$\mathbf{X}\mathbf{w} = \hat{\mathbf{y}}$$

We minimize the sum of squared errors with:

$$\mathbf{w} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

$(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal$ is called the pseudo-inverse.

Inverting $\mathbf{X}^\intercal\mathbf{X}$

But: $\mathbf{X}^\intercal\mathbf{X}$ may be singular.
We have the property: $\mathbf{X}^\intercal\mathbf{X}$ is invertible if $\mathbf{X}$ has linearly independent columns.
This means that: Singular matrices occur with highly correlated features (or duplicate ones).
A way to solve it is to add it a scalar matrix (Hoerl, 1962):

$$\mathbf{w} = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y}$$

Equivalent to adding the term $\lambda||\mathbf{w}||^2$ to the sum of squared errors (SSE).
This process is called a regularization. It is also used in classification.

Regularization

Correlated features result in large values of $\mathbf{w}$ that regularization tries to penalize.
We replace *Loss* with a *Cost*:

$$Cost(\mathbf{w}) = Loss(\mathbf{w}) + \lambda L_q(\mathbf{w}),$$

where

$$L_q = \sum_{i=1}^{n} |w_i|^q.$$

The most frequent regularizations are (ridge regression):

$$L_2 = \sum_{i=1}^{n} w_i^2,$$

and (LASSO regression)

$$L_1 = \sum_{i=1}^{n} |w_i|.$$

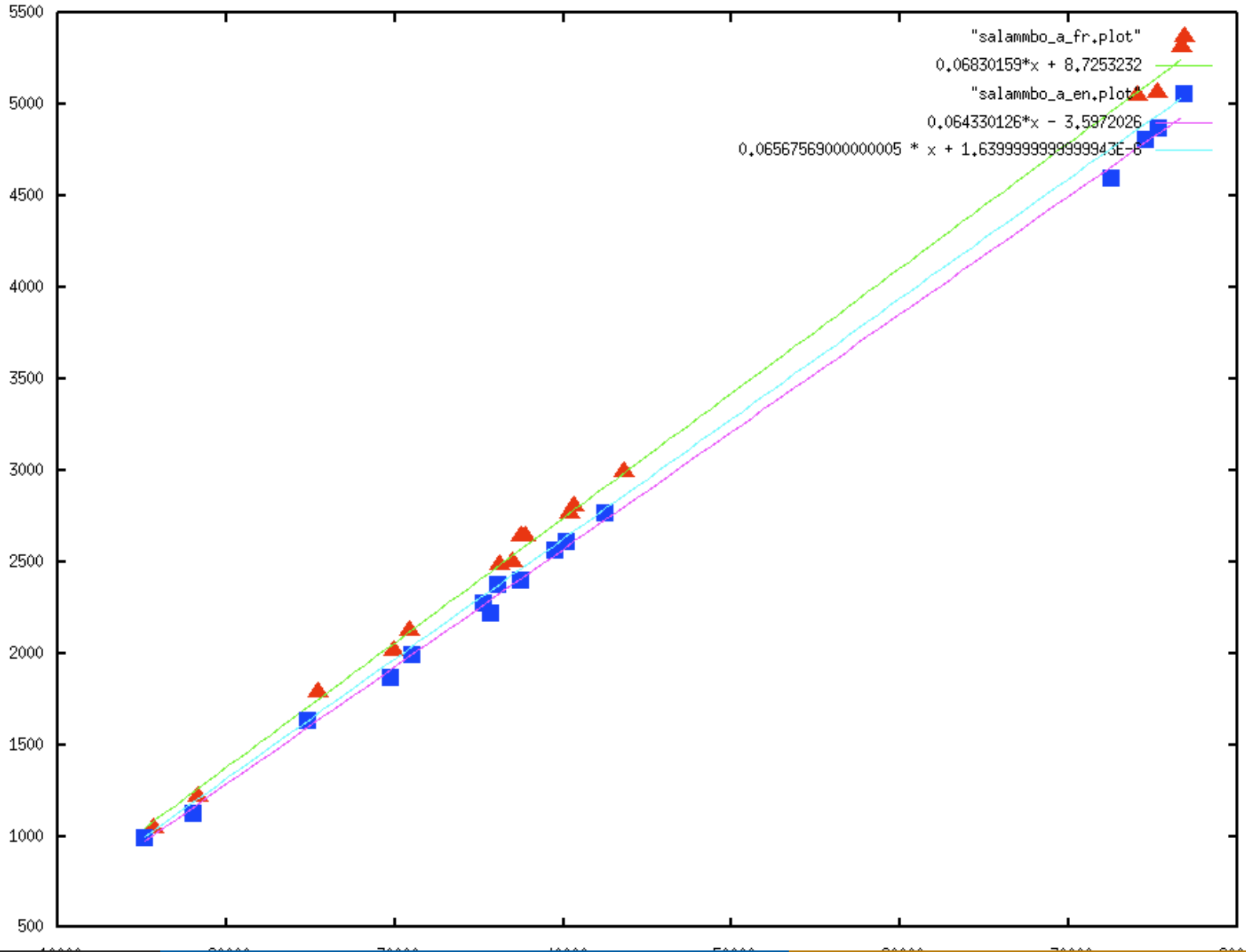
w_0 may be part of the regularization.

Classification vs. Regression

Regression: Given an input, compute a continuous numerical output.
For instance compute the number of *A* occurring in a text in French from the total number of characters.
Having 75,255 characters in Chapter 7, the regression line will predict 5,149 occurrences of *A* (5,062 in reality).

Classification: Given the number of characters and the number of *A* in a text, predict the language: French or English.
For instance, having the pair (75255, 5062), the classifier will predict French.
The classification output is a finite set of values. When there are two values, we have a binary classification.

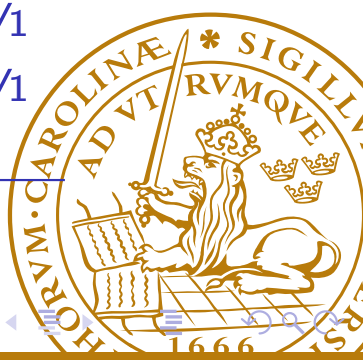
Separating Classes



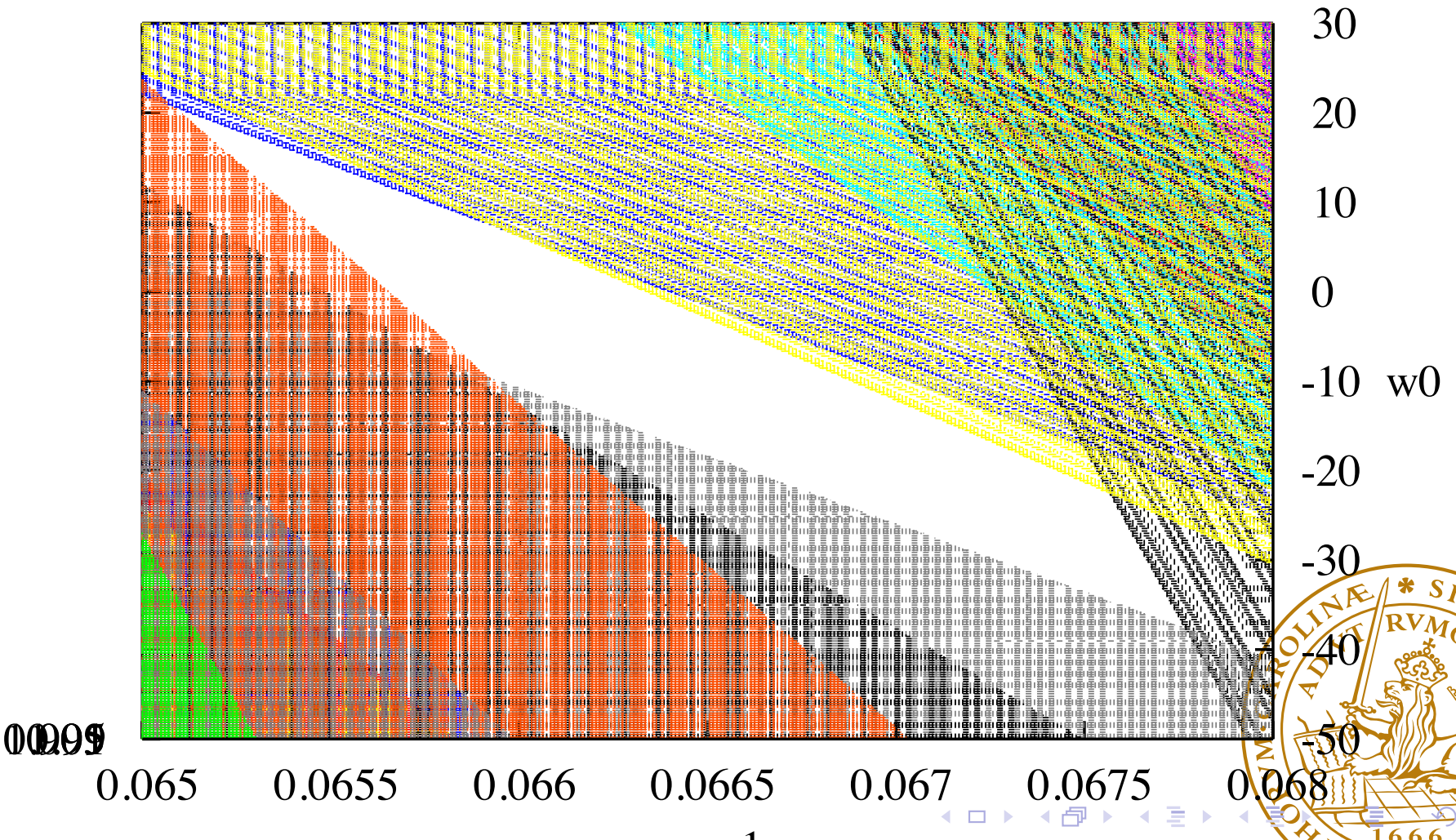
Classification: An Example

$y_i > w_0 + w_1 x_i$ for the set of points: $\{(x_i, y_i) | (x_i, y_i) \in \text{French}\}$ and
$y_i < w_0 + w_1 x_i$ for the set of points: $\{(x_i, y_i) | (x_i, y_i) \in \text{English}\}$,

Chapter	French	English
1	$2503 > w_0 + 36961 w_1$	$2217 < w_0 + 35680 w_1$
2	$2992 > w_0 + 43621 w_1$	$2761 < w_0 + 42514 w_1$
3	$1042 > w_0 + 15694 w_1$	$990 < w_0 + 15162 w_1$
4	$2487 > w_0 + 36231 w_1$	$2274 < w_0 + 35298 w_1$
5	$2014 > w_0 + 29945 w_1$	$1865 < w_0 + 29800 w_1$
6	$2805 > w_0 + 40588 w_1$	$2606 < w_0 + 40255 w_1$
7	$5062 > w_0 + 75255 w_1$	$4805 < w_0 + 74532 w_1$
...		

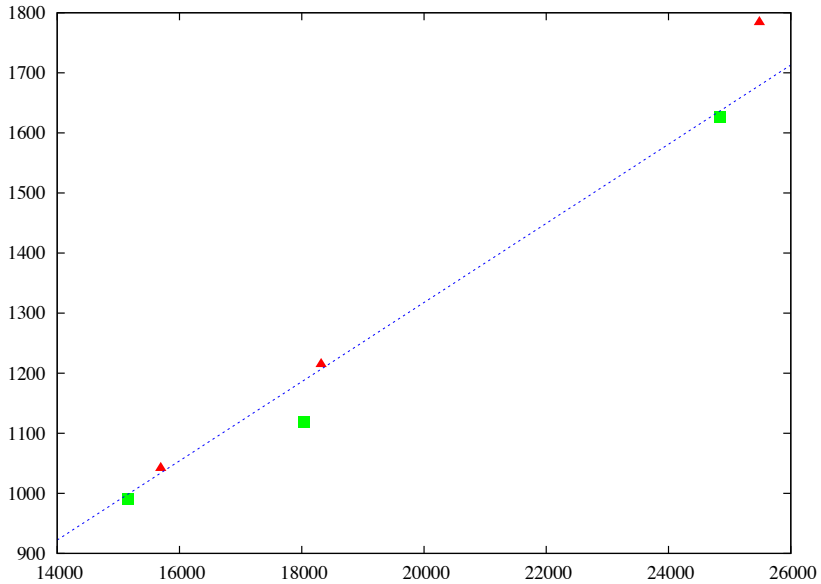


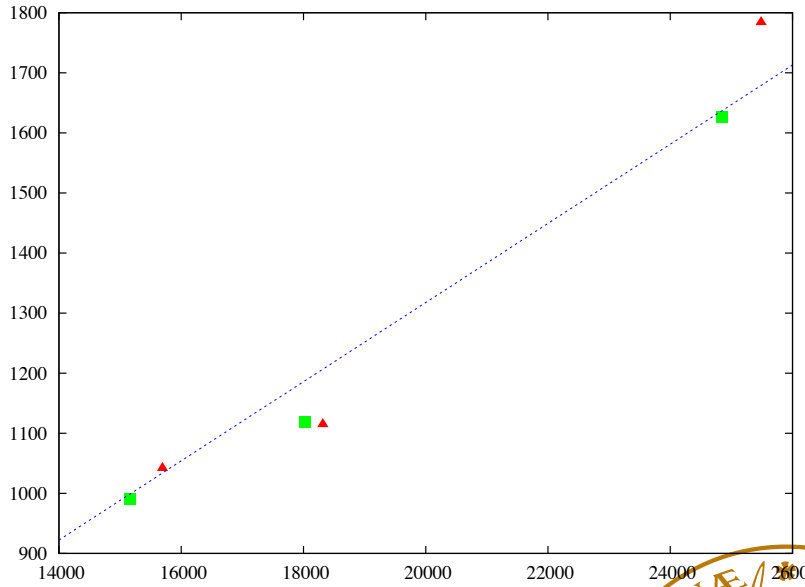
The Inequality System



Separability

Linear classification is based on the linear separability of the classes.
This is not always the case





Dimension of the Feature Space

Regression predicts the value of one of the features given the value of $N-1$ features.
Classification predicts the class given the values of N features.
Compared to regression in our example, the dimension of the vector space used for the classification is $N+1$: The N features and the class.

# characters	# A	Language
36,961	2,503	French
35,680	2,217	English
43,621	2,992	French
42,514	2,761	English
15,694	1,042	French
15,162	990	English
36,231	2,487	French
35,298	2,274	English
etc.		

Classification

We represent classification using a threshold function (a variant of the signum function):

$$H(\mathbf{w}\cdot\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The classification function associates P with 1 and N with 0.
We want to find the separating hyperplane:

$$\begin{aligned} \hat{y}(\mathbf{x}) &= H(\mathbf{w}\cdot\mathbf{x}) \\ &= H(w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n), \end{aligned}$$

given a data set of q examples: $DS = \{(1, x_1^j, x_2^j, ..., x_n^j, y^j) | j : 1..q\}$.
We use $x_0 = 1$ to simplify the equations.
For a binary classifier, y has then two possible values $\{0, 1\}$ corresponding in our example to {French, English}.

Loss Function

To assess the classification, we use the 0/1 loss defined as

$$L_{0/1}(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

The update rule of linear regression is:

$$\begin{aligned} w_i &\leftarrow w_i + \alpha \cdot \sum_{j=1}^{q} x_i^j \cdot (y^j - \hat{y}^j) \\ w_i &\leftarrow w_i + \alpha \cdot \sum_{j=1}^{q} x_i^j \cdot (y^j - (w_0 x_0^j + w_1 x_1^j + w_2 x_2^j + \ldots + w_n x_n^j)) \end{aligned}$$

Using the same idea, we compute the weight updates of a classification as:

$$w_i \leftarrow w_i + \alpha \sum_{j=1}^{q} x_i^j \cdot (y^j - \hat{y}^j)$$

The Perceptron Learning Rule

If all the points are correctly classified, there is no update: $y^j - \hat{y}^j = 0$, either $0-0$ or $1-1$

The loss involves an update only for the misclassified points, either $1-0$ or $0-1$

The update is then: $\alpha \cdot x_i$ or $-\alpha \cdot x_i$

α is generally set to 1 as a division of the weight vector by a constant does not affect the update rule

The perceptron is usually trained one example at a time: stochastic learning

Examples should be selected randomly

In a Two-dimensional Space

Classification with two classes:

$$w_0 + w_1 x_1 + w_2 x_2 > 0$$

$$w_0 + w_1 x_1 + w_2 x_2 < 0,$$

Vectors:

$$\mathbf{x} = (1, x_1, x_2) \quad \text{and} \quad \mathbf{w} = (w_0, w_1, w_2)$$

Stochastic gradient descent. The updates are carried out one example at a time:

$$\begin{aligned} w_0 &\leftarrow w_0 + (y^j - \hat{y}^j) \\ w_1 &\leftarrow w_1 + x_1^j \cdot (y^j - \hat{y}^j) \\ w_2 &\leftarrow w_2 + x_2^j \cdot (y^j - \hat{y}^j) \end{aligned}$$

where $y^j - \hat{y}^j$ is either, 0, -1, or 1.

Stop Conditions

To find a hyperplane, the examples must be separable. This is rarely the case in practice
Workaround:

- Stop learning when the number of misclassified examples is low
- Use $\alpha(t) = \frac{1000}{1000 + t}$ instead of a fixed value. (According to the book)

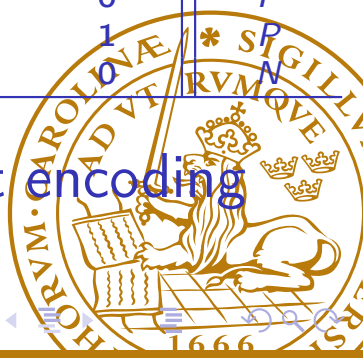
Converting Symbolic Attributes into Numerical Vectors

Linear classifiers are numerical systems.

Symbolic – nominal – attributes are mapped onto vectors of binary values.
A conversion of the weather data set.

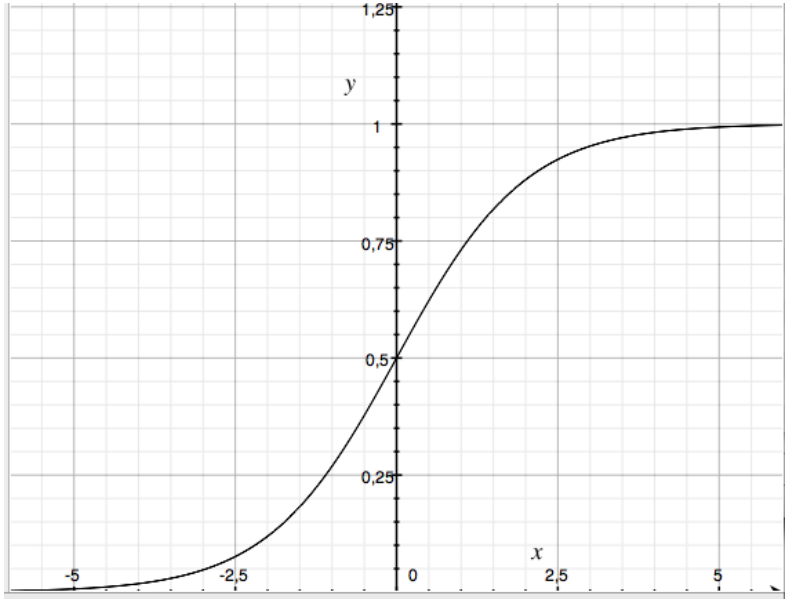
Object	Attributes										Class
	Outlook			Temperature			Humidity		Windy		
	Sunny	Overcast	Rain	Hot	Mild	Cool	High	Normal	True	False	
1	1	0	0	1	0	0	1	0	0	1	*N*
2	1	0	0	1	0	0	1	0	1	0	*N*
3	0	1	0	1	0	0	1	0	0	1	*P*
4	0	0	1	0	1	0	1	0	0	1	*P*
5	0	0	1	0	0	1	0	1	0	1	*P*
6	0	0	1	0	0	1	0	1	1	0	*N*
7	0	1	0	0	0	1	0	1	1	0	*P*
8	1	0	0	0	1	0	1	0	0	1	*N*
9	1	0	0	0	0	1	0	1	0	1	*P*
10	0	0	1	0	1	0	0	1	0	1	*P*
11	1	0	0	0	1	0	0	1	1	0	*P*
12	0	1	0	0	1	0	1	0	1	0	*P*
13	0	1	0	1	0	0	0	1	0	1	*P*
14	0	0	1	0	1	0	1	0	1	0	*N*

This kind of transformation is called contrast coding or one-hot encoding (OHE).



Logistic Regression

The step function is not differentiable; that is why it is often replaced with the logistic curve (Verhulst, 1845, 1848):



$$\begin{aligned} \hat{y}(\mathbf{x}) &= Logistic(\mathbf{w}\cdot\mathbf{x}) \\ &= \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} \end{aligned}$$

Logistic Regression: The Idea

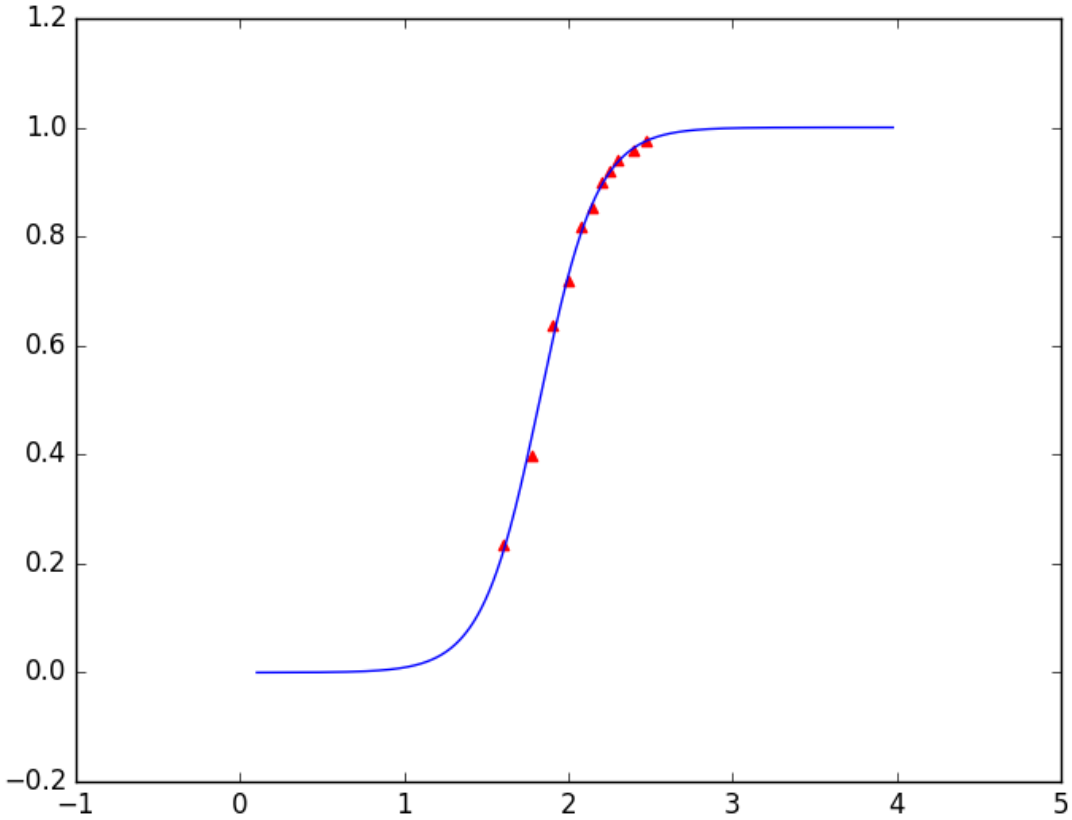
Drug concentration	Number exposed	Survive Class 0	Die Class 1	Mortality rate	Expected mortality
40	462	352	110	.2359	.2206
60	500	301	199	.3980	.4339
80	467	169	298	.6380	.6085
100	515	145	370	.7184	.7291
120	561	102	459	.8182	.8081
140	469	69	400	.8529	.8601
160	550	55	495	.9000	.8952
180	542	43	499	.9207	.9195
200	479	29	450	.9395	.9366
250	497	21	476	.9577	.9624
300	453	11	442	.9757	.9756

Table: A data set. Adapted and simplified from the original article that described how to apply logistic regression to classification by Joseph Berkson, Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association* (1944).

Fitting the Logistic Curve

Berkson fitted the curve to the data using the logarithm of the dose. He obtained:

$$\frac{1}{1+e^{-5.659746x+10.329884}}$$



Logistic Regression

Logistic regression attempts to model the probability of an observation $\mathbf{x}$ to belong to a class:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}$$

and

$$P(y = 0|\mathbf{x}) = \frac{e^{-\mathbf{w}\cdot\mathbf{x}}}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}$$

These probabilities are extremely useful in practice.
The logit assumption (Berkson, 1944) models the odd ratio as a hyperplane:

$$\ln \frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} = \ln \frac{P(y = 1|\mathbf{x})}{1 - P(y = 1|\mathbf{x})} = \mathbf{w} \cdot \mathbf{x}$$

Likelihood of a Classification

Given a data set, DS, containing a partition in two classes, P ($y = 1$) and N ($y = 0$), and a model $\mathbf{w}$, the likelihood to have the classification observed in this data set is:

$$\begin{aligned} L(w) &= \prod_{\mathbf{x}^j \in P} P(y^j = 1|\mathbf{x}^j) \times \prod_{\mathbf{x}^j \in N} P(y^j = 0|\mathbf{x}^j), \\ &= \prod_{\mathbf{x}^j \in P} P(y^j = 1|\mathbf{x}^j) \times \prod_{\mathbf{x}^j \in N} (1 - P(y^j = 1|\mathbf{x}^j)). \end{aligned}$$

We can rewrite the product using y^j as powers of the probabilities as $y^j = 0$, when $\mathbf{x}^j \in P$ and $y^j = 1$, when $\mathbf{x}^j \in N$:

$$\begin{aligned} L(w) &= \prod_{\mathbf{x}^j \in P} P(y^j = 1|\mathbf{x}^j)^{y^j} \times \prod_{\mathbf{x}^j \in N} (1 - P(y^j = 1|\mathbf{x}^j))^{1-y^j}, \\ &= \prod_{(\mathbf{x}^j, y^j) \in DS} P(y^j = 1|\mathbf{x}^j)^{y^j} \times (1 - P(y^j = 1|\mathbf{x}^j))^{1-y^j}. \end{aligned}$$

Maximizing the Likelihood

We train a model by maximizing the likelihood of the observed classification:

$$\hat{w} = \arg\max_{w} \prod_{\mathbf{x}^j \in DS} P(y^j = 1|\mathbf{x}^j)^{y^j} \times (1 - P(y^j = 1|\mathbf{x}^j))^{1-y^j}$$

To maximize this term, it is more convenient to work with sums rather than with products and we take the logarithm of it (log-likelihood):

$$\hat{w} = \arg\max_{w} \sum_{(\mathbf{x}^j, y^j) \in DS} y^j \ln P(y^j = 1|\mathbf{x}^j) + (1 - y^j) \ln(1 - P(y^j = 1|\mathbf{x}^j))$$

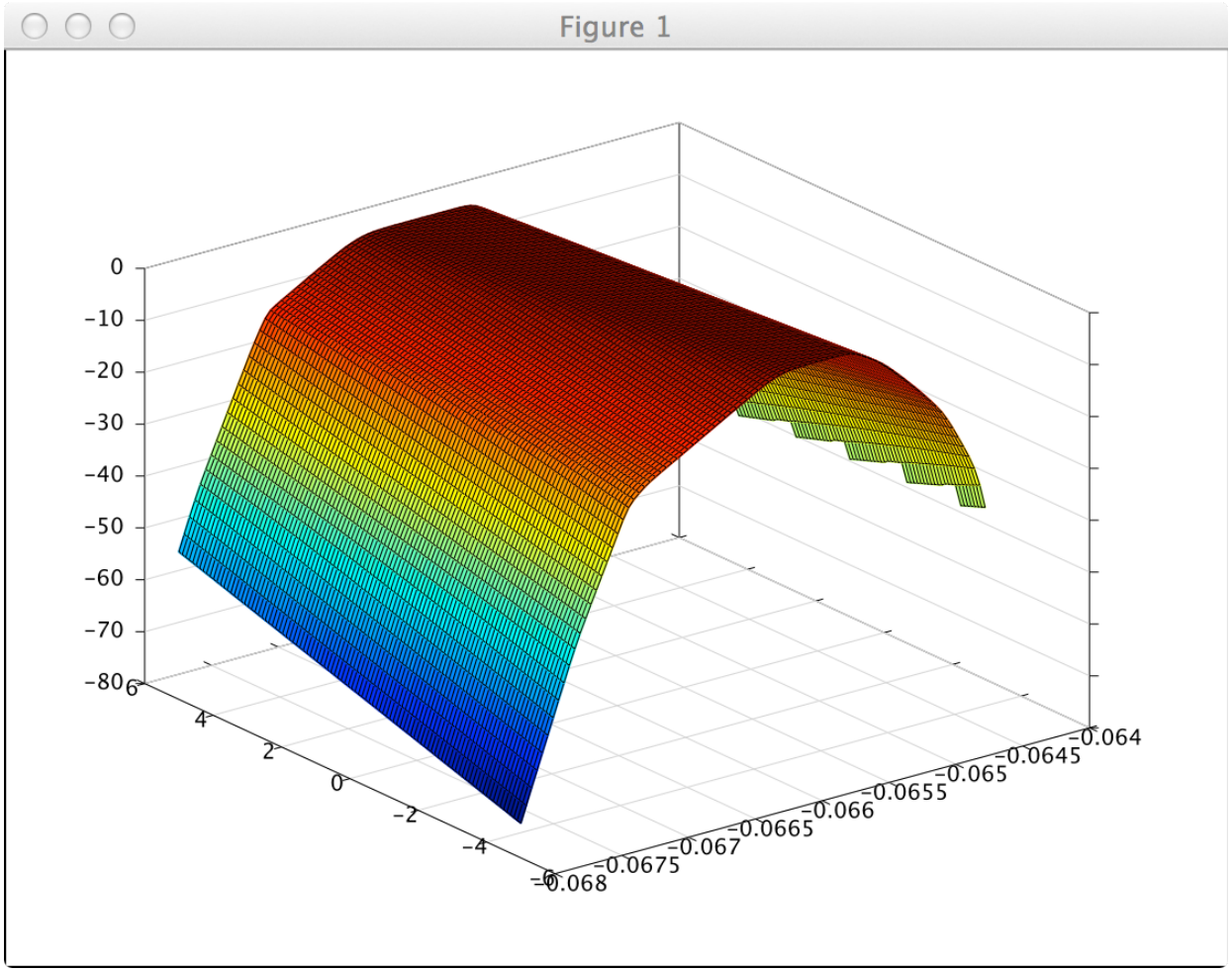
Using the logistic curves to express the probabilities, we have:

$$\hat{w} = \arg\max_{w} \sum_{(\mathbf{x}^j, y^j) \in DS} y^j \ln \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}} + (1 - y^j) \ln \frac{e^{-\mathbf{w}\cdot\mathbf{x}^j}}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}}.$$

In contrast to linear regression that uses least mean squares, here we fit a logistic curve so that it maximizes the likelihood of the classification – partition – observed in the training set.

The Maximum

Graph of the probabilities from the *Salammbô* data set.



The Gradient Ascent

We can use the gradient ascent to compute the maximum.
Using a Taylor expansion: $\ell(\mathbf{w}+\mathbf{v}) = \ell(\mathbf{w}) + \mathbf{v}\cdot\nabla f(\mathbf{w}) + \ldots$
We have then: $\ell(\mathbf{w}+\alpha\nabla\ell(\mathbf{w})) \approx \ell(\mathbf{w}) + \alpha||\nabla\ell(\mathbf{w})||^2$.
The inequality:

$$\ell(\mathbf{w}) < \ell(\mathbf{w}+\alpha\nabla\ell(\mathbf{w}))$$

enables us to move one step up to the maximum.
We then use the iteration:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \alpha_k\nabla\ell(\mathbf{w}_k).$$

Computing the Model

The partial derivatives of the log-likelihood are:

$$\begin{aligned}
\frac{\partial \ell(\mathbf{w})}{\partial w_i} &= \sum_{(\mathbf{x}^j, y^j) \in DS} y^j (1 + e^{-\mathbf{w}.\mathbf{x}^j}) \frac{x_i^j e^{-\mathbf{w}.\mathbf{x}^j}}{(1 + e^{-\mathbf{w}\cdot\mathbf{x}^j})^2} + \\
&\quad (1 - y^j) \frac{1 + e^{-\mathbf{w}.\mathbf{x}^j}}{e^{-\mathbf{w}.\mathbf{x}^j}} \cdot \frac{-x_i^j \cdot e^{-\mathbf{w}\cdot\mathbf{x}^j}(1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}) + x_i^j \cdot e^{-\mathbf{w}.\mathbf{x}^j} \cdot e^{-\mathbf{w}.\mathbf{x}^j}}{(1 + e^{-\mathbf{w}\cdot\mathbf{x}^j})^2}, \\
&= \sum_{(\mathbf{x}^j, y^j) \in DS} y^j \frac{x_i^j e^{-\mathbf{w}.\mathbf{x}^j}}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}} + (1 - y^j) \cdot \frac{-x_i^j \cdot (1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}) + x_i^j \cdot e^{-\mathbf{w}.\mathbf{x}^j}}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}}, \\
&= \sum_{(\mathbf{x}^j, y^j) \in DS} x_i^j \cdot \frac{y^j \cdot (1 + e^{-\mathbf{w}.\mathbf{x}^j}) - 1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}}, \\
&= \sum_{(\mathbf{x}^j, y^j) \in DS} x_i^j \cdot (y^j - \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}^j}}).
\end{aligned}$$

Weight Updates

For $DS = \{(1, x_1^j, x_2^j, ..., x_n^j, y^j) | j : 1..q\}$, the weight updates of logistic regression are then:

- Stochastic gradient ascent:

$$w_i \leftarrow w_i + \alpha \cdot x_i^j \cdot (y^j - \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}^j}})$$

- Batch gradient ascent:

$$w_i \leftarrow w_i + \frac{\alpha}{q} \cdot \sum_{j=1}^{q} x_i^j \cdot (y^j - \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}^j}})$$

As with gradient descent, we stop the iterations when:

$$||\nabla \ell(\mathbf{w})|| < \varepsilon.$$

Multinomial (Multiclass) Logistic Regression

Generalizing logistic regression from

$$\ln \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \mathbf{w} \cdot \mathbf{x}$$

to a multiclass setting is easy.
Using $y = 0$ as a pivot, we have for C classes:

$$\begin{aligned}
\ln \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} &= \mathbf{w}_1 \cdot \mathbf{x} \\
\ln \frac{P(y=2|\mathbf{x})}{P(y=0|\mathbf{x})} &= \mathbf{w}_2 \cdot \mathbf{x} \\
\dots & \\
\ln \frac{P(y=C-1|\mathbf{x})}{P(y=0|\mathbf{x})} &= \mathbf{w}_{C-1} \cdot \mathbf{x}
\end{aligned}$$

and

$$\sum_{k=0}^{C-1} P(y=k|\mathbf{x}) = 1.$$

Multinomial (Multiclass) Logistic Regression (II)

We obtain:

$$P(y=0|\mathbf{x}) = \frac{1}{1+\sum_{i=1}^{C-1} e^{\mathbf{w}_i \cdot \mathbf{x}}}$$

$$P(y=k|\mathbf{x}) = \frac{e^{\mathbf{w}_k \cdot \mathbf{x}}}{1+\sum_{i=1}^{C-1} e^{\mathbf{w}_i \cdot \mathbf{x}}}, k = 1, 2, ..., C-1.$$

Equivalent to:

$$P(y=k|\mathbf{x}) = \frac{e^{\mathbf{w}_k \cdot \mathbf{x}}}{\sum_{i=0}^{C-1} e^{\mathbf{w}_i \cdot \mathbf{x}}}, k = 0, 1, 2, ..., C-1.$$

This function is often called the *softmax* function.

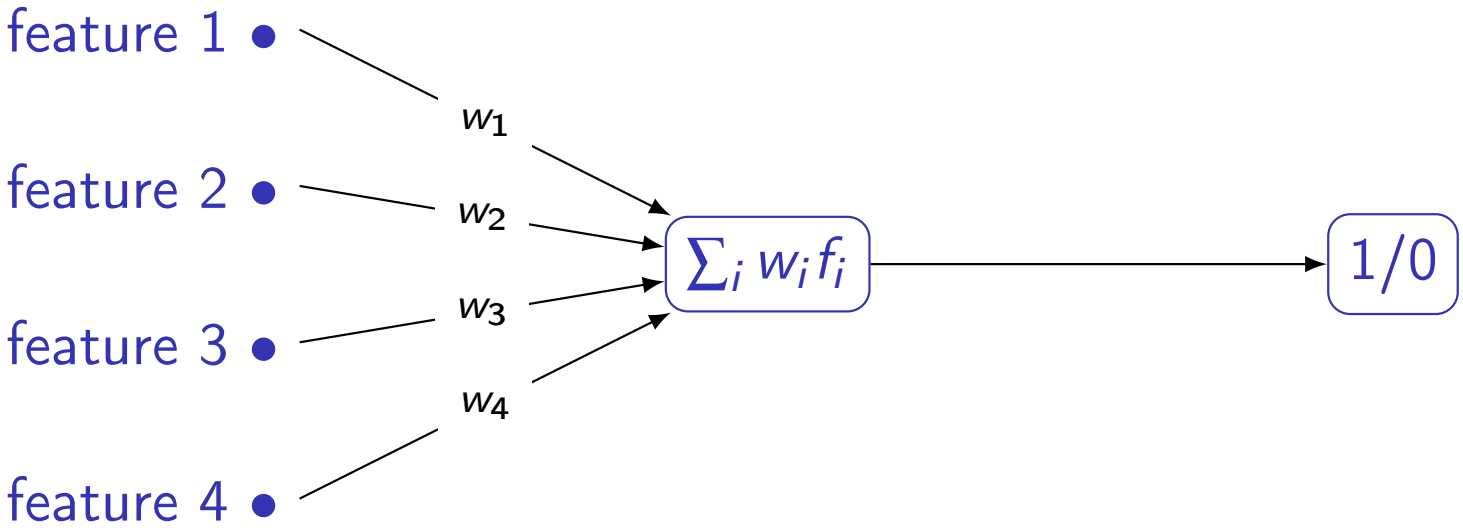
Neural Networks

Learning devices loosely inspired by brain neuron
Artificial neural networks abundantly use physiological metaphors, such as synapses, cortex, etc.
In our context, can be reframed as a multistage logistic regression to handle nonlinearities
Training is complex and demanding
Fell out of favor in the 1990-2000s and experienced a revival in the name of deep learning
Available programming environments include, Keras (TensorFlow and Theano), PaddlePaddle, and Caffe:

- https://www.tensorflow.org/
- http://deeplearning.net/tutorial/
- http://deeplearning.net/software/theano/
- https://keras.io/
- http://www.paddlepaddle.org/
- http://caffe.berkeleyvision.org/

Neural Networks: Representation

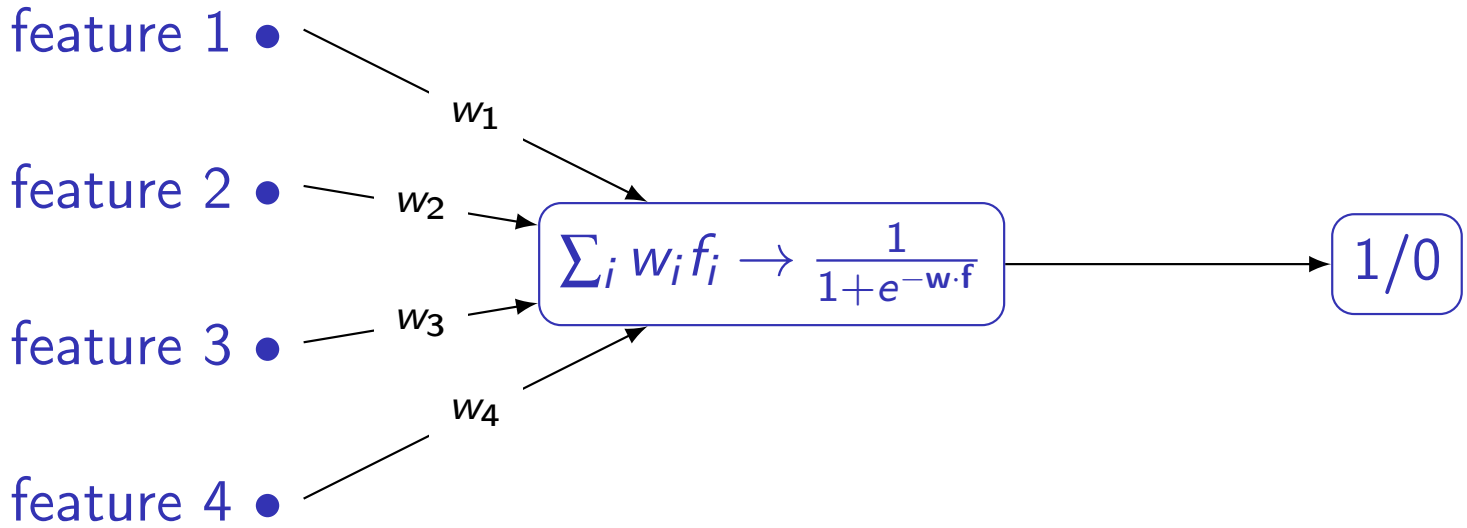
Another representation of the perceptron:



The base network: An input layer and an output layer

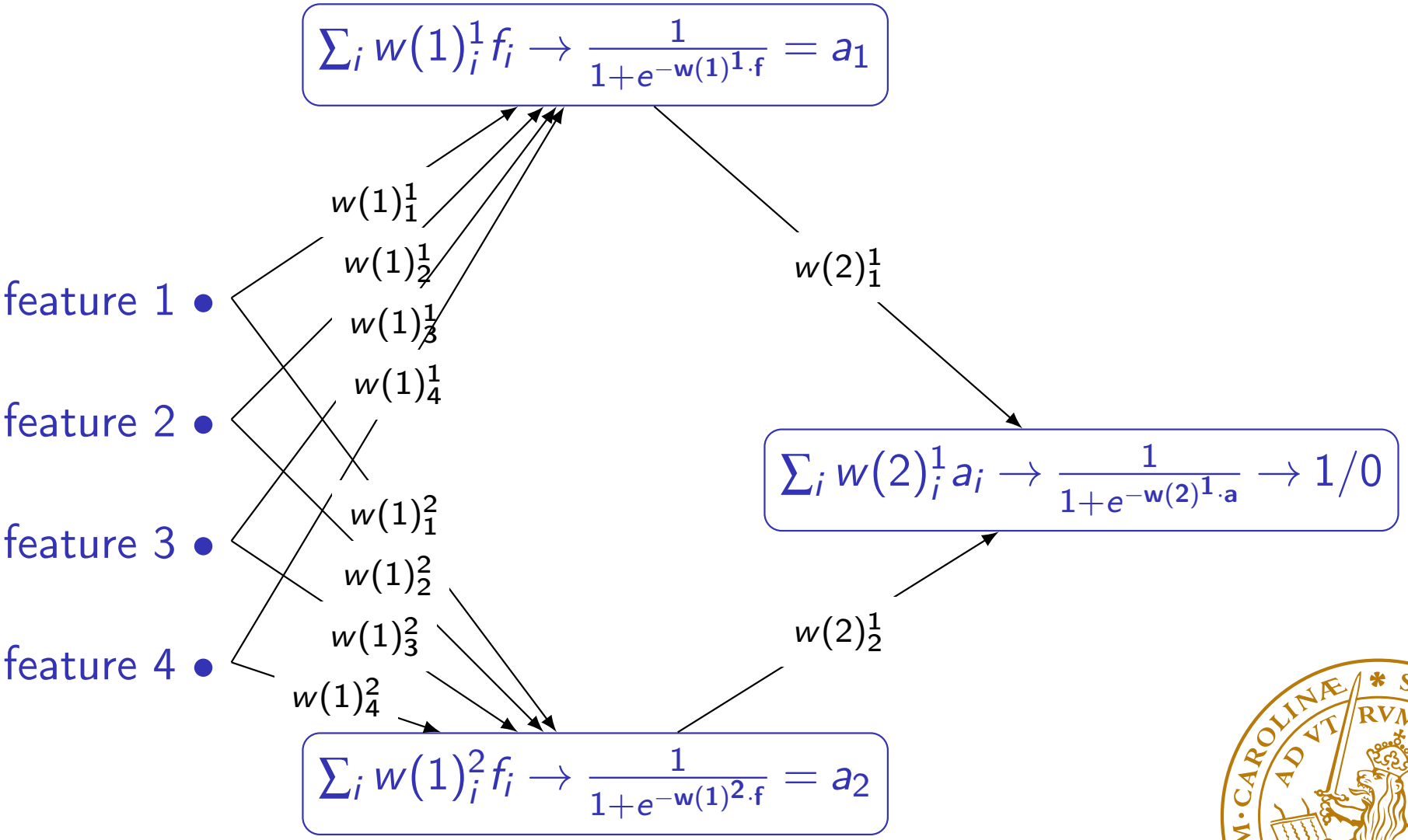
Neural Networks: Activation Function

And logistic regression:

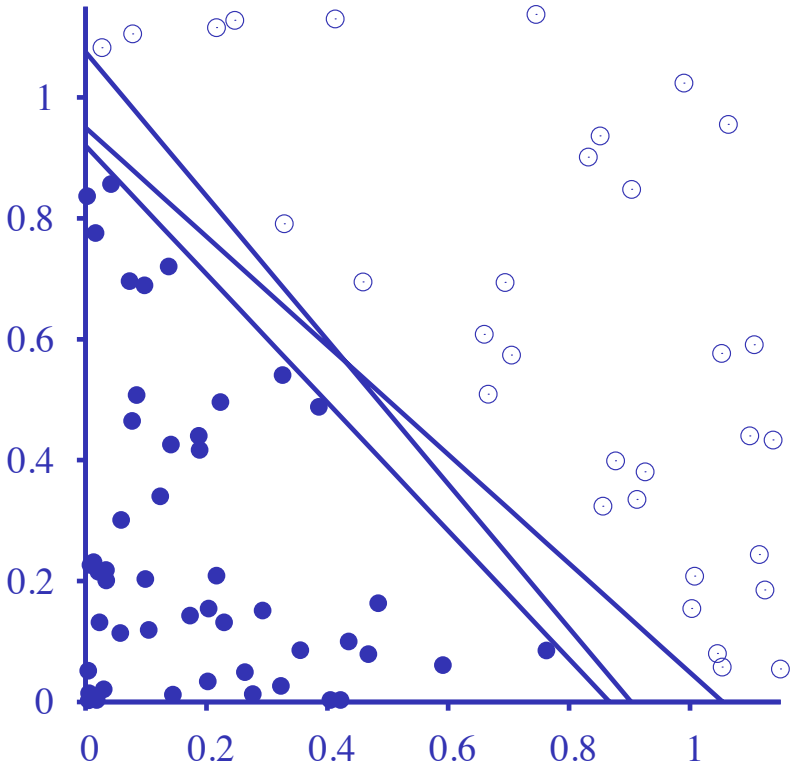


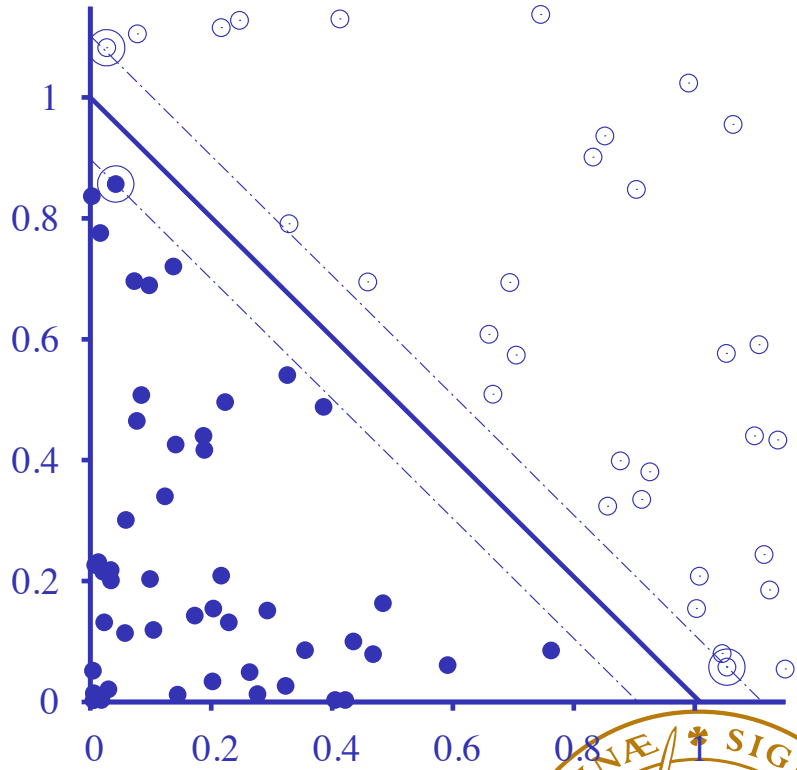
The logistic function is the activation function of the node

Neural Networks: Hidden Layers



Support Vector Machines





From the textbook: Stuart Russell and Peter Norvig, *Artificial Intelligence*, 3rd ed., 2010, page 745.

Support Vector Machines (II)

Support vector machines (SVM) maximize the margin between the separating hyperplane and the examples
The hyperplane is set at equal distance between the closest points of each class
The closest points are called the support vectors
Support vector machines can handle nonseparable examples using a soft margin or kernels.

Problem Description

With two classes, support vector machines use the notation: $y = +1$ or $y = -1$.
We know from geometry and vector analysis that the distance of a point A to a hyperplane Hyp defined by the equation:

$$w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n = 0$$

is given by the formula:

$$d(A, Hyp) = \frac{|w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n|}{\sqrt{w_1^2 + w_2^2 + ... + w_n^2}}.$$

We want to find the hyperplane maximizing b so that:

$$y^j \cdot \frac{w_0 + \mathbf{w} \cdot \mathbf{x^j}}{||\mathbf{w}||} \geqslant b$$

for all the points in the data set and with $\mathbf{w} = (w_1, ..., w_n)$.

Maximizing the Margin

We normalize the equation with $b = \frac{1}{||\mathbf{w}||}$.
To find the maximal margin b, we minimize $||\mathbf{w}||^2$ with the constraint:

$$y^j \cdot (w_0 + \mathbf{w} \cdot \mathbf{x^j}) \geqslant 1,$$

where y^j values are either +1 or -1 depending on the class of the example.
We can solve this using gradient descent and hinge loss defined as:

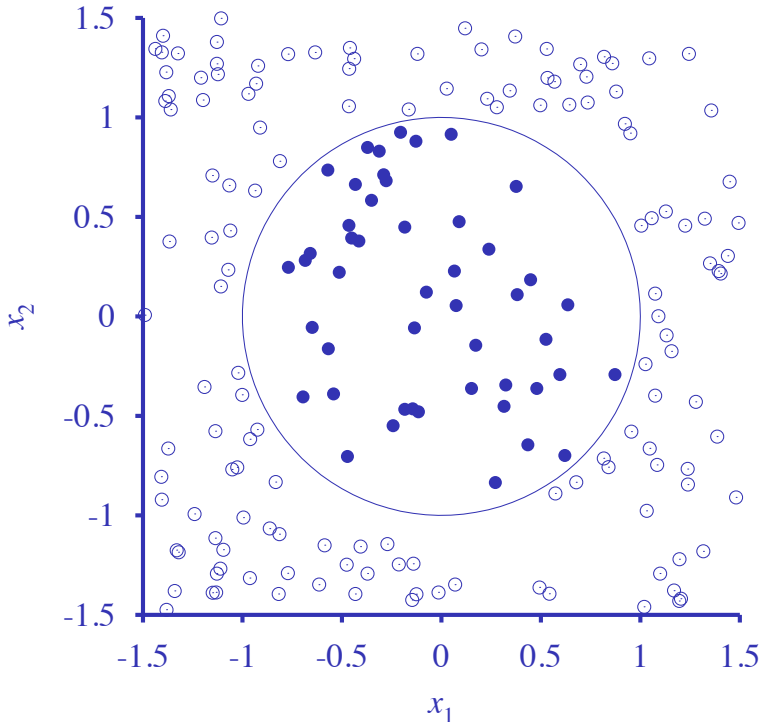
$$\max\{0, 1 - y\mathbf{w}^T\mathbf{x}\}, \quad y \in \{-1, +1\},$$

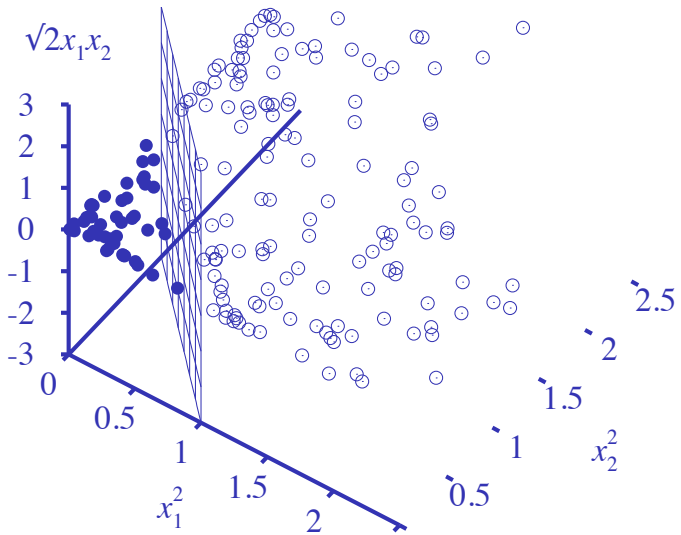
where the gradient is defined as:

$$\nabla L_{hinge} = \begin{cases} -y \cdot \mathbf{x} & \text{if } y\mathbf{w}^T\mathbf{x} < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Kernels

Many data sets are not linearly separable as in left-hand side figure.





From the textbook: Stuart Russell and Peter Norvig, *Artificial Intelligence*, 3rd ed., 2010, page 747.

It is always possible to find a space of higher dimension, where the points will be separable, here 3.

The Kernel Trick

If we map the input data (x_1, x_2) to $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$, we produce linearly separable classes:

$$F(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

We replace $\mathbf{x_j x_t}$ with $F(\mathbf{x_j})F(\mathbf{x_t})$ in the Lagrangian:

$$L(\alpha) = \sum_{j=1}^{q} \alpha_j - \frac{1}{2}\sum_{j,t}^{q} \alpha_j \alpha_t y^j y^t F(\mathbf{x_j})F(\mathbf{x_t}).$$

In the case of F, we have

$$\begin{aligned} F(\mathbf{x}) \cdot F(\mathbf{z}) &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 \\ &= (\mathbf{x} \cdot \mathbf{z})^2 \end{aligned}$$

In many cases, we can replace $\mathbf{x_j} \cdot \mathbf{x_t}$ with a kernel function $K(\mathbf{x_j}, \mathbf{x_t})$ like $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2$

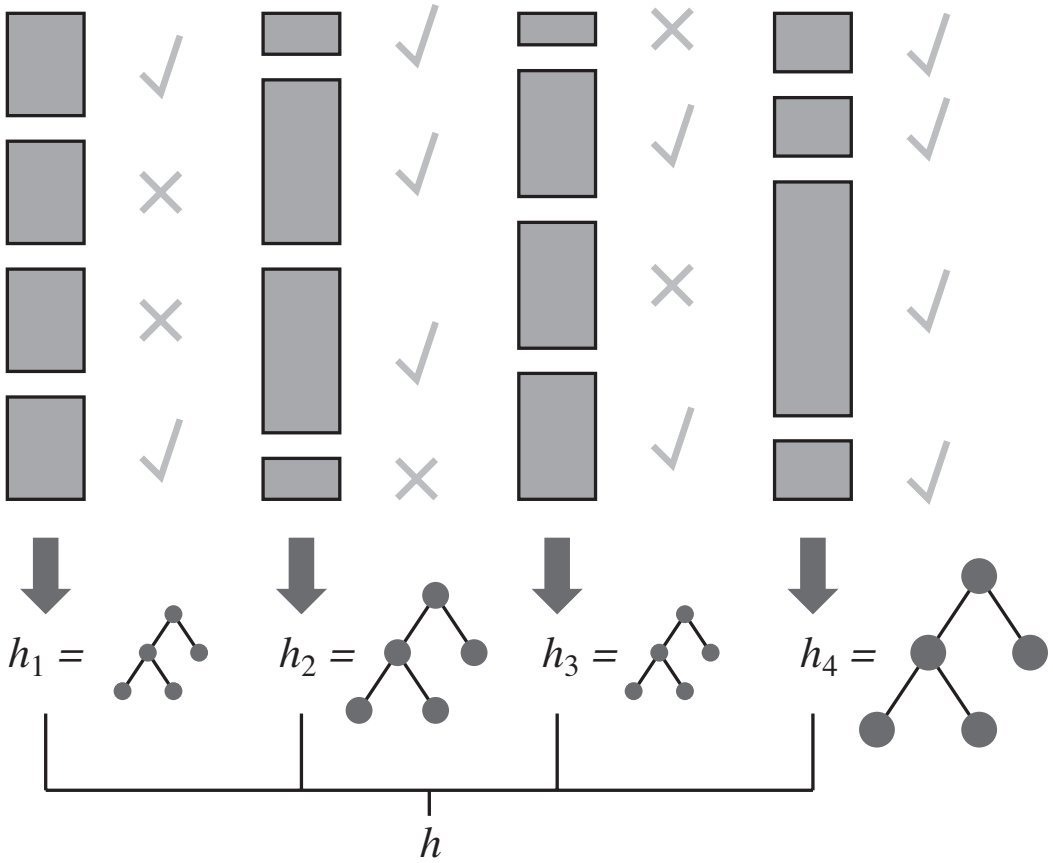
Ensemble Learning

Idea: Use many classifiers instead of one
Boosting is an example of ensemble learning
Uses a set of K weak classifiers and gives a weight to the examples: $w(i)$
This means that example i counts as $w(i)$ examples
Decision stumps (decision trees with one level) are popular weak classifiers
The idea of the AdaBoost algorithm is:

1. Learn a first stump on the training set. Each example has a weight of 1
2. Set a higher weight to misclassified examples (think of just duplicating them in the data set)
3. Learn a second decision stump.
4. and so on until we have K decision stumps

The final classifier is a weighted combination of the K classifiers, where the weights are computed from the performance of each individual classifier.

Ensemble Learning



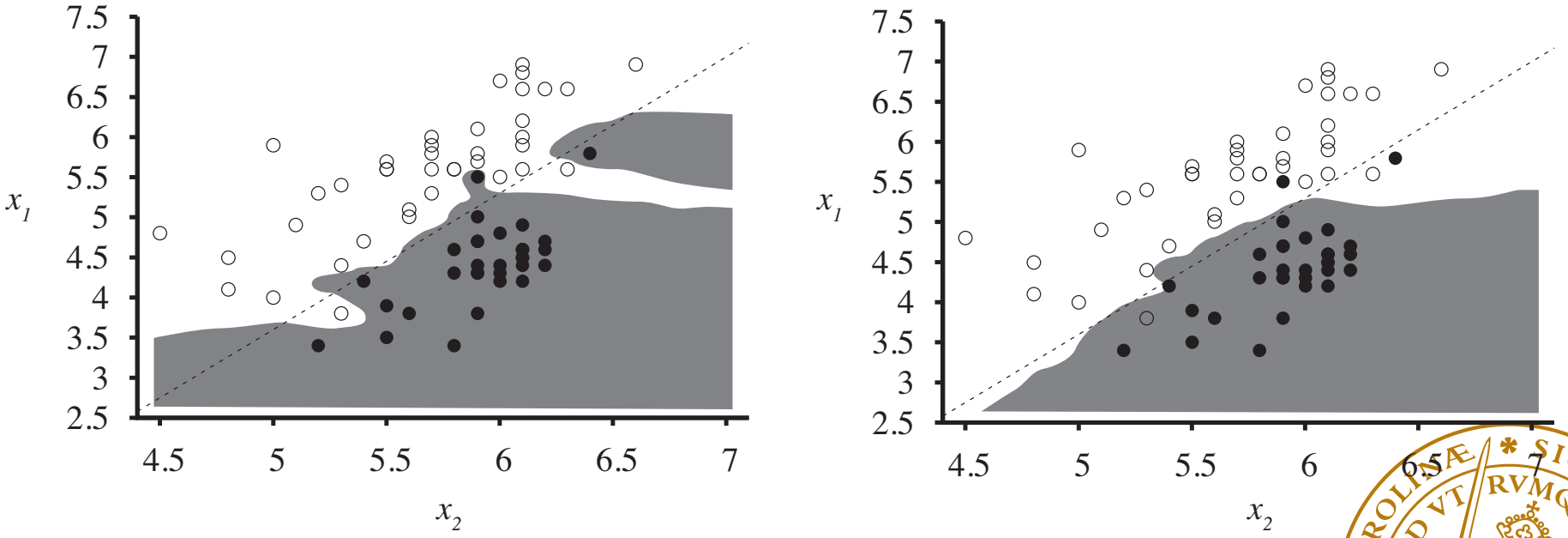
From the textbook: Stuart Russell and Peter Norvig, *Artificial Intelligence*, 3rd ed., 2010, page 750.

Nonparametric Models

When there is no parametric model, such as a hyperplane, we can try to search the k nearest neighbors of $\mathbf{x}$: $NN(k, \mathbf{x})$

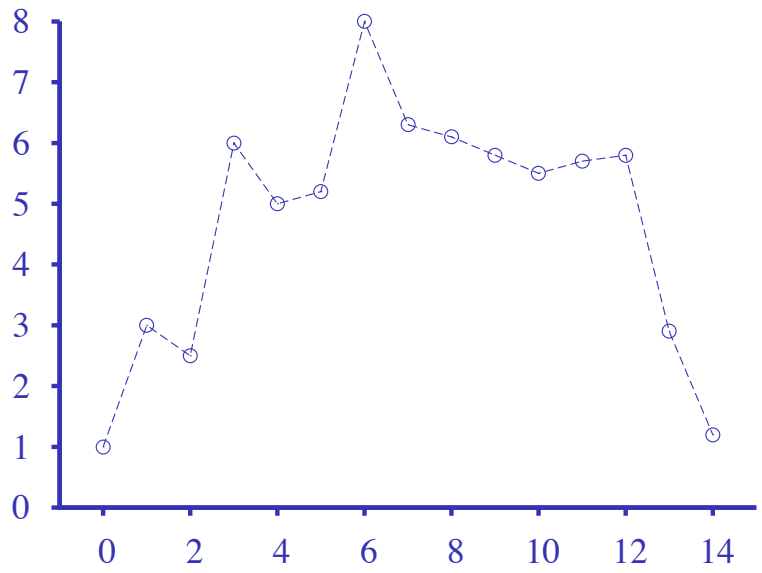
Given a position in the space, we decide on a class by looking at the neighbor's class

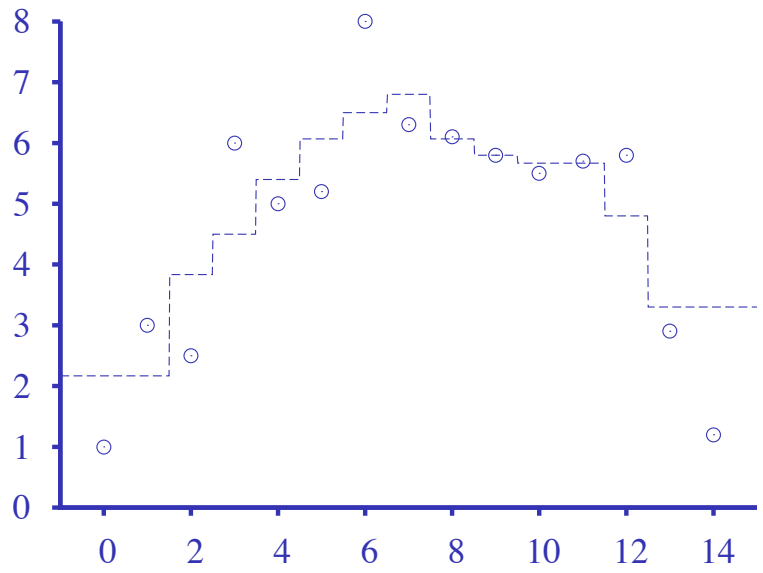
The distance is given by: $L^p(\mathbf{x}, \mathbf{y}) = (\sum_i |x_i - y_i|^p)^{\frac{1}{p}}$

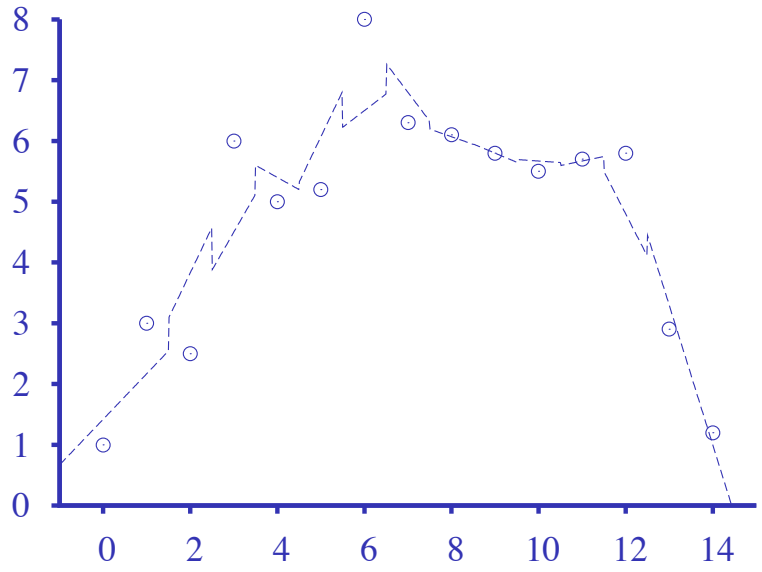


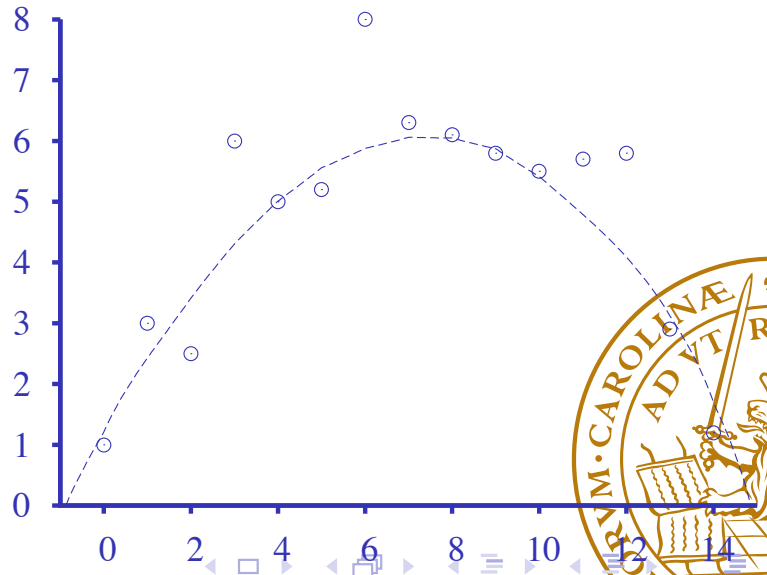
k nearest neighbors with $k = 1$ and $k = 5$. From the textbook: Stuart Russell and Peter Norvig, *Artificial Intelligence*, 3rd ed., 2010, page 738.

Nonparametric Regression









scikit-learn

scikit-learn is a popular and comprehensive machine–learning library in Python built on top of numpy.
The classifiers use two main functions: `fit()` to train a model and `predict()` to predict a class. In the scikit-learn documentation, the functions adopt a notation, where:

- $\mathbf{x}$ denotes a feature vector (the predictors) describing one observation and $\mathbf{X}$, a feature matrix representing the dataset;
- y denotes the class (or response or target) of one observation and $\mathbf{y}$, the class vector for the whole dataset.

scikit-learn: Data

```
import numpy as np

y_train = np.array(
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
     1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])

X_train = np.array(
    [[35680, 2217], [42514, 2761], [15162, 990], [35298, 2274],
     [29800, 1865], [40255, 2606], [74532, 4805], [37464, 2396],
     [31030, 1993], [24843, 1627], [36172, 2375], [39552, 2560],
     [72545, 4597], [75352, 4871], [18031, 1119], [36961, 2503],
     [43621, 2992], [15694, 1042], [36231, 2487], [29945, 2014],
     [40588, 2805], [75255, 5062], [37709, 2643], [30899, 2126],
     [25486, 1784], [37497, 2641], [40398, 2766], [74105, 5047],
     [76725, 5312], [18317, 1215]
     ])
```

scikit-learn: Training a Model

We then select and fit a model, here logistic regression with default parameters, with the lines:

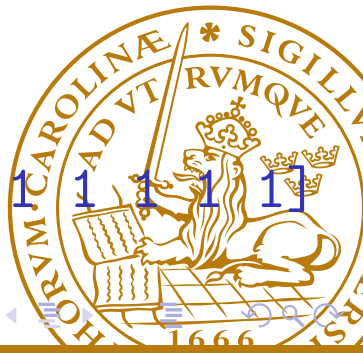
```
from sklearn import linear_model

classifier = linear_model.LogisticRegression()
model = classifier.fit(X_train, y_train)
```

Once the model is trained, we can apply it to new observations. The next instruction just reapplies it to the training set:

```
y_test_predicted = classifier.predict(X_train)
```

which predicts exactly the classes we had in this set:

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```



scikit-learn: Crossvalidation

scikit-learn has built-in cross validation functions. The code below shows an example of it with a 5-fold cross validation and a score corresponding to the accuracy:

```
from sklearn.model_selection import cross_val_score

scores = cross_val_score(classifier, X_train, y_train, cv=5,
                         scoring='accuracy')
print('Score', scores.mean())
```

scikit-learn: Nominal Data

$$\mathbf{X} = \begin{bmatrix} \text{Sunny} & \text{Hot} & \text{High} & \text{False} \\ \text{Sunny} & \text{Hot} & \text{High} & \text{True} \\ \text{Overcast} & \text{Hot} & \text{High} & \text{False} \\ \text{Rain} & \text{Mild} & \text{High} & \text{False} \\ \text{Rain} & \text{Cool} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Cool} & \text{Normal} & \text{True} \\ \text{Overcast} & \text{Cool} & \text{Normal} & \text{True} \\ \text{Sunny} & \text{Mild} & \text{High} & \text{False} \\ \text{Sunny} & \text{Cool} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Mild} & \text{Normal} & \text{False} \\ \text{Sunny} & \text{Mild} & \text{Normal} & \text{True} \\ \text{Overcast} & \text{Mild} & \text{High} & \text{True} \\ \text{Overcast} & \text{Hot} & \text{Normal} & \text{False} \\ \text{Rain} & \text{Mild} & \text{High} & \text{True} \end{bmatrix}; \mathbf{y} = \begin{bmatrix} \text{N} \\ \text{N} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{N} \\ \text{P} \\ \text{N} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{P} \\ \text{N} \end{bmatrix}$$

scikit-learn: Vectorizing the Data

To convert the data to numbers, store **X** in a list of dictionaries and use `DictVectorizer`:

```
y = [0 if symb == 'no' else 1 for symb in y_symbols]
```

and we vectorize the features with `DictVectorizer`:

```
from sklearn.feature_extraction import DictVectorizer

vec = DictVectorizer(sparse=False) # Should be true
X = vec.fit_transform(X_dict)
```

scikit-learn: Data Vectorization

```
array([[ 1.,  0.,  0.,  0.,  1.,  0.,  1.,  0.,  1.,  0.],
       [ 1.,  0.,  0.,  0.,  1.,  0.,  1.,  0.,  0.,  1.],
       [ 1.,  0.,  1.,  0.,  0.,  0.,  1.,  0.,  1.,  0.],
       [ 1.,  0.,  0.,  1.,  0.,  0.,  0.,  1.,  1.,  0.],
       [ 0.,  1.,  0.,  1.,  0.,  1.,  0.,  0.,  1.,  0.],
       [ 0.,  1.,  0.,  1.,  0.,  1.,  0.,  0.,  0.,  1.],
       [ 0.,  1.,  1.,  0.,  0.,  1.,  0.,  0.,  0.,  1.],
       [ 1.,  0.,  0.,  0.,  1.,  0.,  0.,  1.,  1.,  0.],
       [ 0.,  1.,  0.,  0.,  1.,  1.,  0.,  0.,  1.,  0.],
       [ 0.,  1.,  0.,  1.,  0.,  0.,  0.,  1.,  1.,  0.],
       [ 0.,  1.,  0.,  0.,  1.,  0.,  0.,  1.,  0.,  1.],
       [ 1.,  0.,  1.,  0.,  0.,  0.,  0.,  1.,  0.,  1.],
       [ 0.,  1.,  1.,  0.,  0.,  0.,  1.,  0.,  1.,  0.],
       [ 1.,  0.,  0.,  1.,  0.,  0.,  0.,  1.,  0.,  1.]])
```

Further Resources

A fine videolecture: Peter Norvig, *Statistical Learning as the Ultimate Agile Development Tool*
http://videolectures.net/cikm08_norvig_slatuad/

An interesting paper: Pedro Domingos, A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, 55 (10), 78-87, 2012

Online courses: openclassroom.stanford.edu, udacity.com, coursera.org, edx.org, etc.

Books: Two excellent books to go further: *An Introduction to Statistical Learning with Applications in R* (start with this one) and *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*
http://web.stanford.edu/~hastie/pub.htm

Tips from the expert: An interview with #1 on Kaggle:
http://blog.kaggle.com/2015/11/09/profiling-top-kagglers-gilberto-titericz-new-1-in-t