UNIVERSITÀ DEGLI STUDI
DI TRENTO

HANDOUTS

“The Sum of the Parts is other than the whole”

PATRIZIO BELLAN

FRANCESCO MANTEGNA

Center for Mind/Brain Sciences
UNIVERSITY OF TRENTO - Italy
CIMeC

Index:

Bayes' Theorem

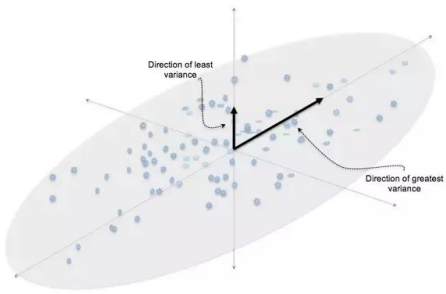
Bayes' rule is a rigorous method for interpreting evidence in the context of previous experience or knowledge.

$$\Pr(\mathrm{A}|\mathrm{X}) = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A}) + \Pr(\mathrm{X}|\text{not A})\Pr(\text{not A})}$$

Decision Tree

The classification technique is a systematic approach to build classification models from input data. The decision tree classifiers organize a series of test questions and conditions in a tree structure. So, a decision tree is a graph that uses a branching method to illustrate every possible decision outcome.
The root and internal nodes contain attribute test conditions that are used to separate records showing different characteristics. To each terminal node a class label is assigned. During the test phase, starting from the root node, we apply the test condition to the input and follow the appropriate branch based on the test outcome. When it reaches the leaf node, the class label associated with the leaf node is then assigned to the input.

Singular Value Decomposition



The singular value decomposition (SVD) is a factorization of a real or complex matrix.

The directions along which there is greatest variance are referred to as the "principal components".

Geometrically, the SVD means that spheres of the proper dimension in the domain are transformed into ellipsoids in the codomain. Since the transformation may not be injective, the dimension of the ellipsoid is at most the dimension of the sphere. So you get some distortion along some axes and some collapsing along other axes.

Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components

When do we use SVD and when PCA?

PCA is used for finding the directions where most of the energy of the vectors lies, whereas SVD is a factorization of matrix into ortho-normal spaces.
There is a very direct mathematical relation between SVD (Singular Value Decomposition) and PCA (Principal Component Analysis).
But, even if the two algorithms deliver essentially the same result (a set of "new axes" constructed from linear combinations of the original feature space axes in which the dataset is plotted), and formally both solutions can be used to calculate the same principal components and their correspondent to singular values, the extra step of calculating the covariance matrix can lead to numerical rounding errors when calculating the vectors.

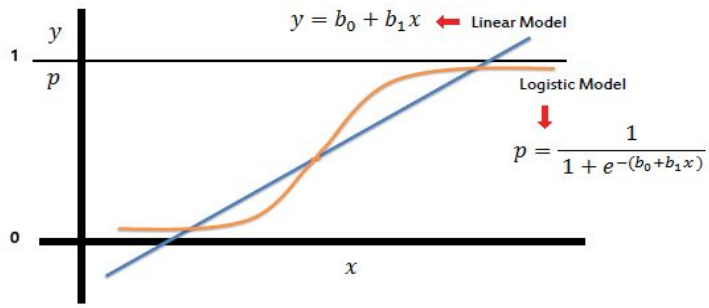
Logistic Regression

The logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The task is to estimate the log odds of an event. Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

In linear regression we try to predict the value of $y(i)$ for the i'th example $x(i)$ using a linear function $y=h\theta(x)$.

The function $\sigma(z)$ is an S-shaped function that "squashes" the value of x into the range [0,1], so that we may interpret $h\theta(x)$ as a probability. The probability $P(y=1|x)=h\theta(x)$ is large when x belongs to the "1" class and



small when *x* belongs to the “0” class.

Linear Regression

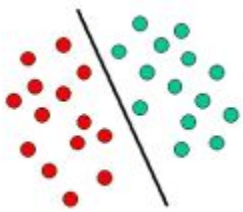
Linear regression uses the general linear equation $Y=b0+\sum(biXi)+\epsilon Y$, where *Y* is a dependent variable and independent variables *Xi*. Both Y and X are continuous variables. The term ϵ is the variance that is not explained by the model and is usually just called "error".

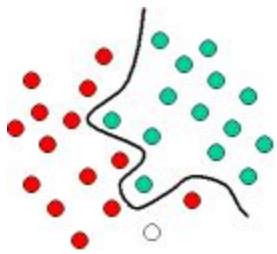
What is the difference between linear regression and logistic regression?

Logistic regression is a generalized linear model (GLM) procedure using the same basic formula of Linear Regression. However, instead of the continuous *YY*, it is regressing for the probability of a categorical outcome.

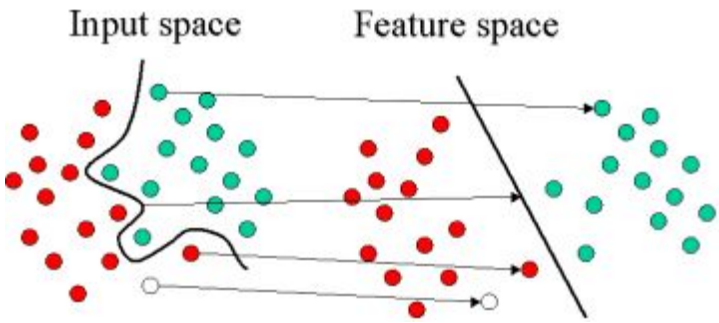
Support Vector Machine

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.





Most classification tasks, however, are not that simple , and often more complex structures are needed in order to make an optimal separation (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks. Data are mapped (rearranged) using a set of mathematical functions, known as kernels. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks.



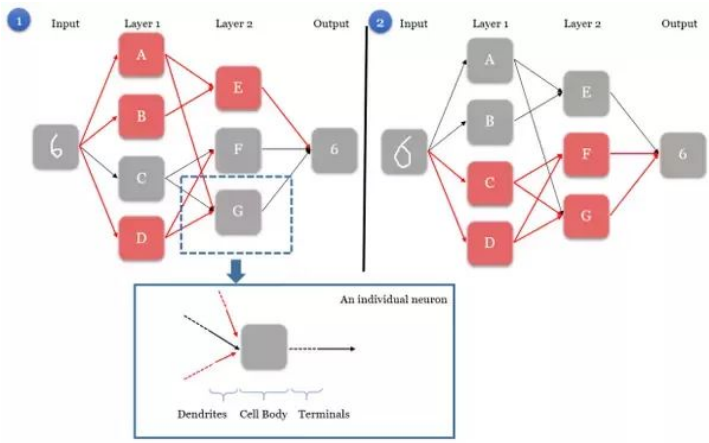
Kernel functions

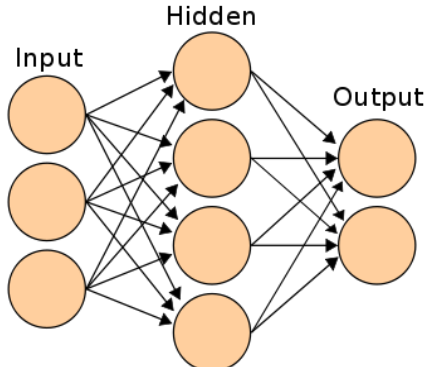
$$K(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear} \\ (\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C)^d & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X}_i - \mathbf{X}_j|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C) & \text{Sigmoid} \end{cases}$$

The kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ. *Gamma* is an adjustable parameter of certain kernel functions. The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Artificial Neural Network

Artificial neural networks (ANNs) is a computational model able to be any real function. It does that taking the biological model as example. ANNs simply map inputs to outputs. They do so by adding biases to the inputs as well as multiplying those inputs with weights. Those weights and biases together are known as a model, which is responsible for the guesses the ANN makes about inputs. The weights and biases can and will be adjusted, depending on how accurate they are with their guesses. Using a loss function, we measure how wrong each guess is by contrasting it with a ground-truth answer. Then we take the error produced by the loss function, and we use it to adjust the weights and biases. Each neuron makes a single decision by a weighted approach. The inputs are weighted because some inputs are more important than others.





Two main factors are involved during the training of the model:

- A metric to evaluate the model's accuracy
- Rules that govern whether neurons are activated or not

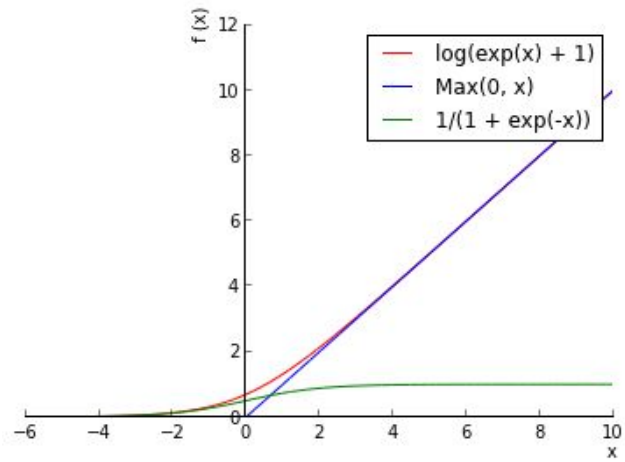
A common metric to evaluate model accuracy is the sum of the squared errors (SSE).

The ANN model will try to minimize the loss by changing its internal weight coefficients.

Rectified linear unit (ReLU) Function

The ReLU is an activation function used in the hidden layers of the network. It is commonly used in classification problems because introduce sparsity among neurons and does not face the gradient vanishing problems (present in the other activation function such sigmoid and tahn.

$$f(x)=\sum_{i=1}^{inf}\sigma(x-i+0.5)\approx log(1+e^x)$$



Other common activation function are:

Sigmoid unit : $f(x)=\frac{1}{1+exp(-x)}$

Tanh unit: $f(x)=tanh(x)$

Artificial Neural Networks: Why do we use the softmax function for the output layer?

Simply, sigmoid could work as well but softmax do the job better. The sigmoid function in the final layer works just in case the output admits multiple "true" answers (so, if it is not a probability distribution). Classification problems can take advantage of the fact that the classes are mutually exclusive. Mathematically, using a SoftMax activation function is basically equivalent to use a Logistic Regression over the features extracted from the layer before the final Fully Connected layer. Softmax activation is basically the normalized exponential probability of class observations represented as neuron activations.

UNIVERSITÀ DEGLI STUDI DI TRENTO

Our Results

	Accuracy
Bayes	0.888
Decision Tree	0.9004
Logistic Regression (regul.=1)	**0.9146**
Logistic Regression (regul.=10)	0.9074
Logistic Regression (regul.=100)	0.9036
Logistic Regression (regul.=1000)	0.9004

	Linear Kernel	rbf Kernel
regul = 1 Gamma = 1	**0.83220**	0.845949
regul = 1 Gamma = 10	0.83220	0.846673
regul = 1 Gamma = 100	0.83220	0.847878
regul = 10 Gamma = 1	0.83220	0.846190
regul = 10 Gamma = 10	0.83220	0.848360
regul = 10 Gamma = 100	0.83220	**0.849083**

	Linear Kernel	rbf Kernel
regul = 100 Gamma = 1	0.83220	0.8457087
regul = 100 Gamma = 10	0.83220	0.8488428
regul = 100 Gamma = 100	0.83220	0.8488428
regul = 1000 Gamma = 1	0.83220	0.8478784
regul = 1000 Gamma = 10	0.83220	0.8481195
regul = 1000 Gamma = 100	0.83220	0.8459498

	1 hidden layer	**2 hidden layer**	**3 hidden layer**
Accuracy	**0.967213**	0.951543	0.844021
Neurons first layer	500	250	250
Neurons second layer	0	125	125
Neurons third layer	0	0	25
starting learning rate	1e-02	e-02	e-02