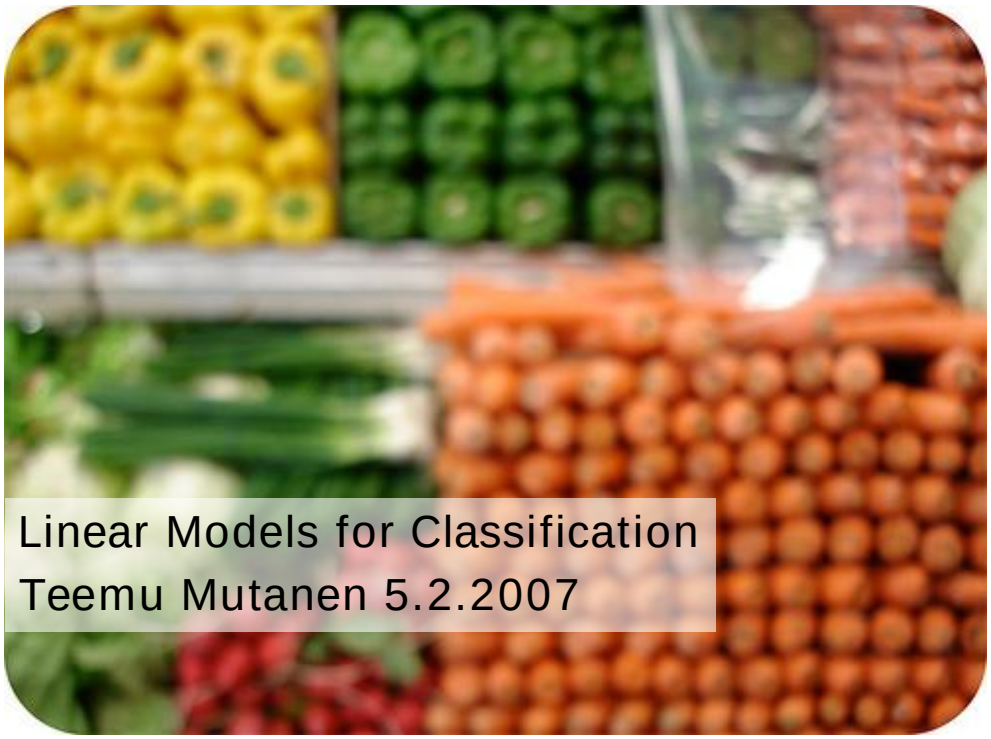


The presentation is based on the C. Bishop's book Pattern Recognition and Machine Learning (ISBN: 0387310738), Chapter 4.

The presentation can be found online at:
http://www.cis.hut.fi/Opinnot/T-61.6020/

We are going to present you today a brief introduction to the are of classification, and we will also discuss linear methods. Antti Sorjamaa will be joining me today for presenting some interesting insights on this subject. My name is Teemu Mutanen. Good afternoon.

First I would like to be clear about how I understand the classification problem and what it means as we choose the proper method to deal with it.

Classification problem

There is basically only one type of classification problem: You have whole bunch of data and then you want to chop it into pieces. But being as intelligent as you are, you want to do the choping intelligently…This brings me to the two pictures you see here.

The picture on the left shows you a mix of vegetable. There are only two types of vegetables present: carrots and beans. Each class has its own characteristics (orange/green, short and thick vs long and thin). While the picture on the right shows you the problem of how to eat citrus fruit. This has been solved by knife, and the person who did the work had somehow acquired information about the inner formation of an orange. I like this picture because the person with the knife has tried to follow the natural patterns of orange slices and these slices as result are somewhat similar to the natural slices.

The classification to multiple classes basically covers also the area of two classes, but the reason why I want to point out the difference is that usually with two classes, the problem is prediction of other class assignments. If you are allergic to something you know what I mean... In this carrots and beans classification, the missclassifications of carrrots could be serious, the beans don’t matter. Seriously, you can make no mistake with false classificaiton of carrots... and this is usually the case when two class classification is carried out: the assignments to one class matters.

Why?

Predictive power

Aid to communication

Failures might be interesting

Simple - Easy

I would like to point out couple interesting characteristics why you should consider classification on your problem… And more, why you should choose linear models to help you.

1. A good linear classification model has predictive power. Each new intance may be classified to a class based on the classification model. Although the assignment have some uncertainty, the predictions are useful and helpful.

2. The classification helps communication. Look at the orange slices, TV-viers were classified as such in the 80's. I am not sure about right now, maybe? They assign you to one class, and then they target some specific programs during the prime time they hope you watch the TV…

3. Failures of classification models may point out interesting behavior or special objects that deserve attention. This is actually known as outlier detection, a field where the goal is to find special objects - outliers - which don't fit in the normal behavior. To best of my understanding this is one of the things in counter terrorism.

4. The linear classification models are simple to formulate. And easy to understand. The effect and significanse of each attribute? Can you make it more simple than this. Do you really need some fuzzy logic, it's still carrots and beans… Or five hidden layers, those are still carrots and beans!

Linear discriminant function

Decision surfaces are hyperplanes

input: vector x
output: assignments

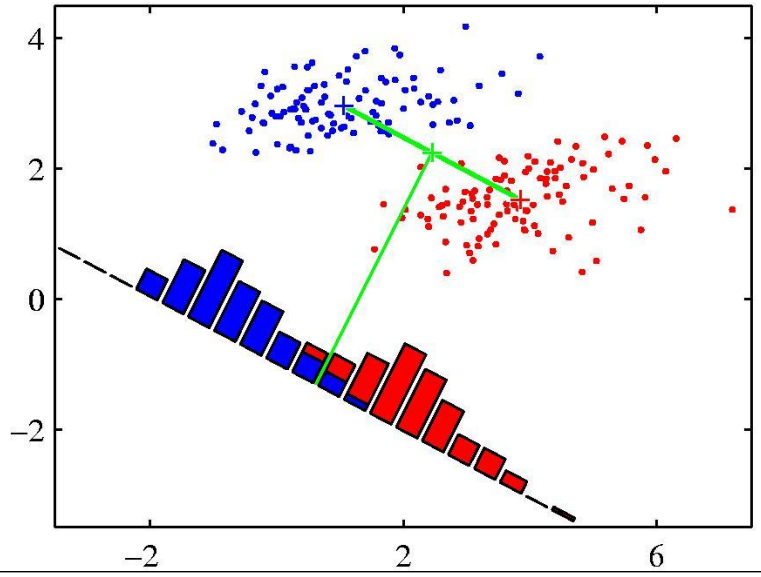
Disjoint classes – one class and only one class

Illustration of the geometry of a linear discriminant function in two dimension. Bishop's book figure 4.1

The least squares approach doesn't succeed here as we can recall that it corresponds to maximum likelihood under the assumption of a Gaussian conditional distribution, whereas binary target vectors clearly have a distribution that is far from Gaussian.

I will discuss next two ideas as alternative nonprobabilistic methods for setting the parameters in the linear classification models. These are Fisher's linear discriminant and the perceptron algorithm.

Fisher's linear discriminant



One way to view a linear classification model is in terms of dimensionality reduction. Cosider now this D-dimensional input vector x which is projected down to one dimension. In general this means a considerable loss of information, the classes that are well separated in D-dimensions could strongly overlap in one dimension.

The picture you see here is an example of how the components of the weight vector w has been adjusted. The picture shows the case when the samples are projected onto the line joining the means.

This type of problem arises from the strongly nondiagonal covariances of the class distributions.

Fisher's idea for avoiding this is as follows: maximize a function that will give a large separation between the projected class means while also giving a small variance within each class.

Fisher's linear discriminant

Large separation between the projected class means while also giving a small variance within each class.

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

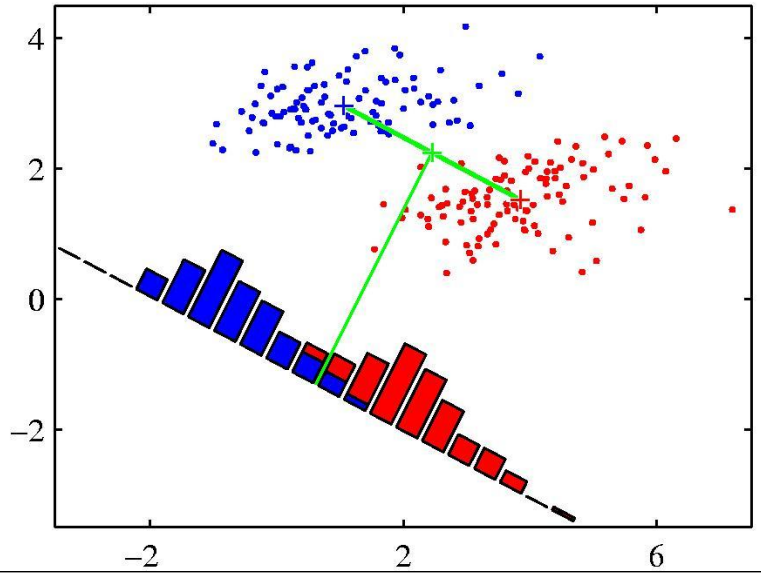
Fisher's criterion is defined to be the ratio of the between-class variance to the within-class variance.

The actual Fisher's criterion is obtained by differentiating this respect to w and thus J(w) is maximized when…

(Equation 4.30.)

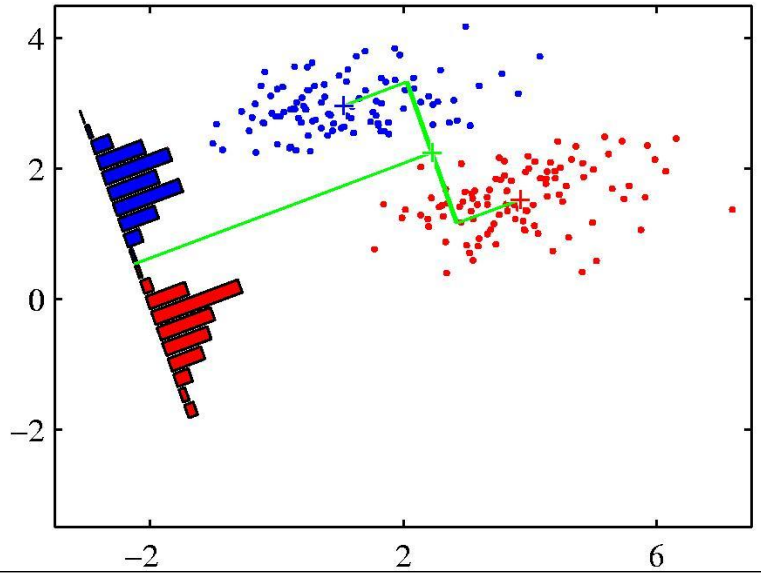
The criterion is not actually discriminant but rather a specific choise of direction for projection of the data down to one dimension.

Fisher's linear discriminant



Now we are back with this example of the projected dimension. This next plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

Fisher's linear discriminant



The Perceptron Algorithm

Frank Rosenblatt (1962)

Two class model

Input vector x is transformed to give a feature vector $\phi(x)$

$$y(x) = f(w^T \phi(x))$$

The Perceptron Algorithm

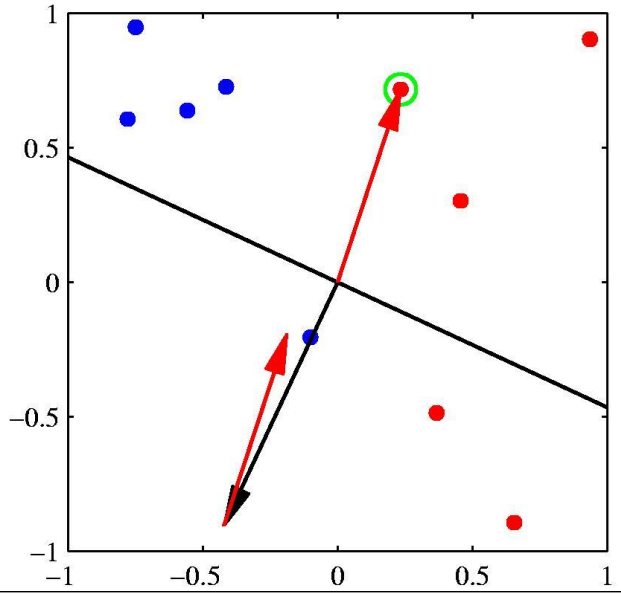
Missclassified patterns

$$0 \quad \text{or} \quad -w^T \phi(x_n) t_n$$

The perceptron criterion:

$$E_p(w) = -\sum_{n \in \mathrm{M}} w^T \phi_n t_n$$

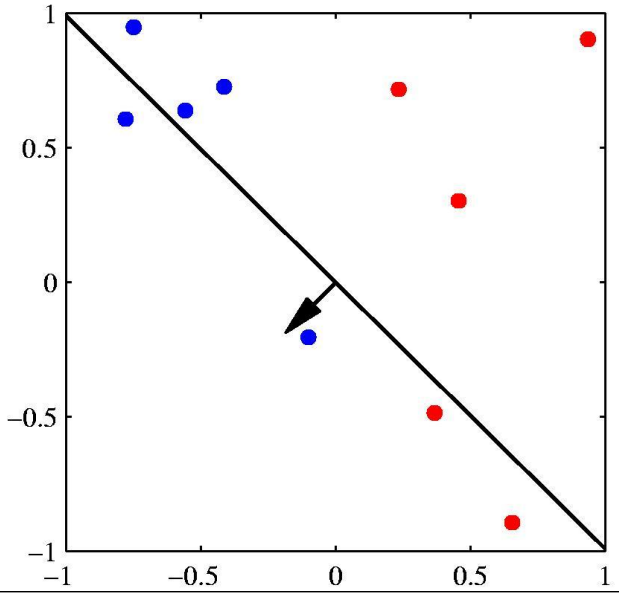
The Perceptron Algorithm



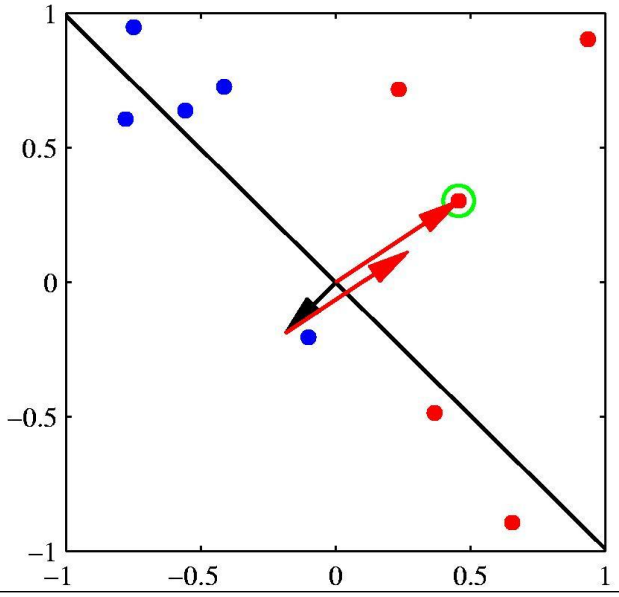
The perceptron learning algorothm has a simple interpretation, as follows. We cycle through the training patterns in turn, and evaluate the perceptron function.

If the pattern is correcly classified, then weigh vector remains unchanged, whereas if missclassfied, then for class 1 we add the vector phi onto the current estimate weight vector W, and for class 2 we substract it.

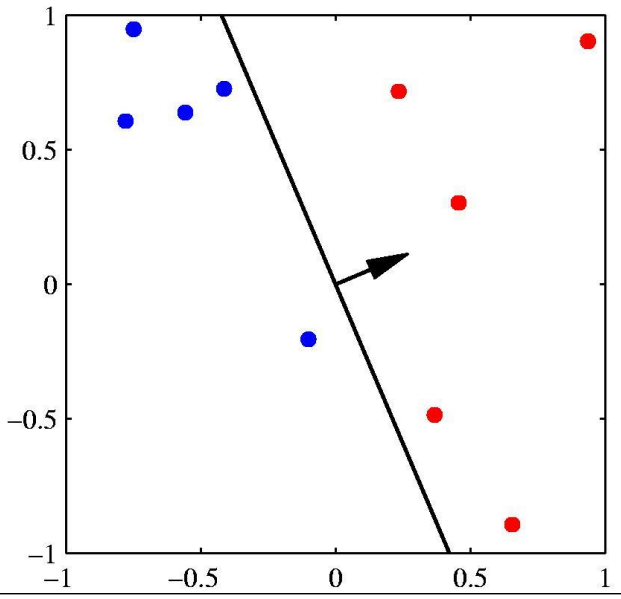
The Perceptron Algorithm



The Perceptron Algorithm



The Perceptron Algorithm



Let me go through that again. Replay!

Perceptron convergence theoren states that if the data is linearly separable then the perceptron learning algorithm is guaranteed to find an exact solution in finite number of steps. Before termination there is no indication whether the data is separable. And also if the data is not linearly separable then the perceptron learning algorithm will never converge.

There is couple clear difficulties in perceptron algorithm:
The perceptron doesn't provide probabilistic outputs
It doesn't generalize readily to K>2 situations
And it is based on linear combinations of fixed basis functions.

Probabilistic Generative Models

Class-conditional densities $p(x|C_k)$

Class priors $p(C_k)$

Posterior probabilities $p(C_k|x)$

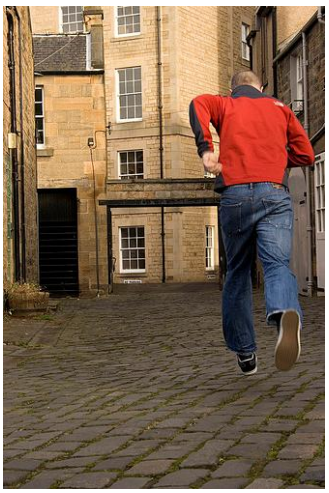
Two class case:

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)}$$

Take the money and run!

Consider the following exercises:

4.9*, 4.11**, and 4.16*



Yes, I will end here.

Antti will start from here, and I think he will present atleast some interesting insights about logistic regression.

But before that I want to point out couple interesting exercises. You might have already done some of them, but if you haven't I recommend you to do so. These three exercises covers key terminology of this chapter.

Thank you for your attention!

photos © winding road, greentheory, dirtyone30, and colinj @ flickr

Thank you very much.

The original photo used as the front page image is available online at:
http://www.flickr.com/photos/greentheory/344342777/in/pool-photojojo/

The pictures in the slide 2 are available online at:
http://www.flickr.com/photos/mariasariego/357683476/in/pool-photojojo/
http://www.flickr.com/photos/colinj/19511628/

And on slide 11:
http://www.flickr.com/photos/dirtyone30/375499861/in/pool-photojojo/

Other figures are originally published in the book by C. Bishop,
the figures can be found at:
http://research.microsoft.com/~cmbishop/PRML/webfigs.htm

For contact: teemu.mutanen@vtt.fi
Questions and comments are always welcome.