Modeling Dependence of Daily Stock Prices and Making Predictions of Future Movements

Taavi Tamkivi, prof Tõnu Kollo

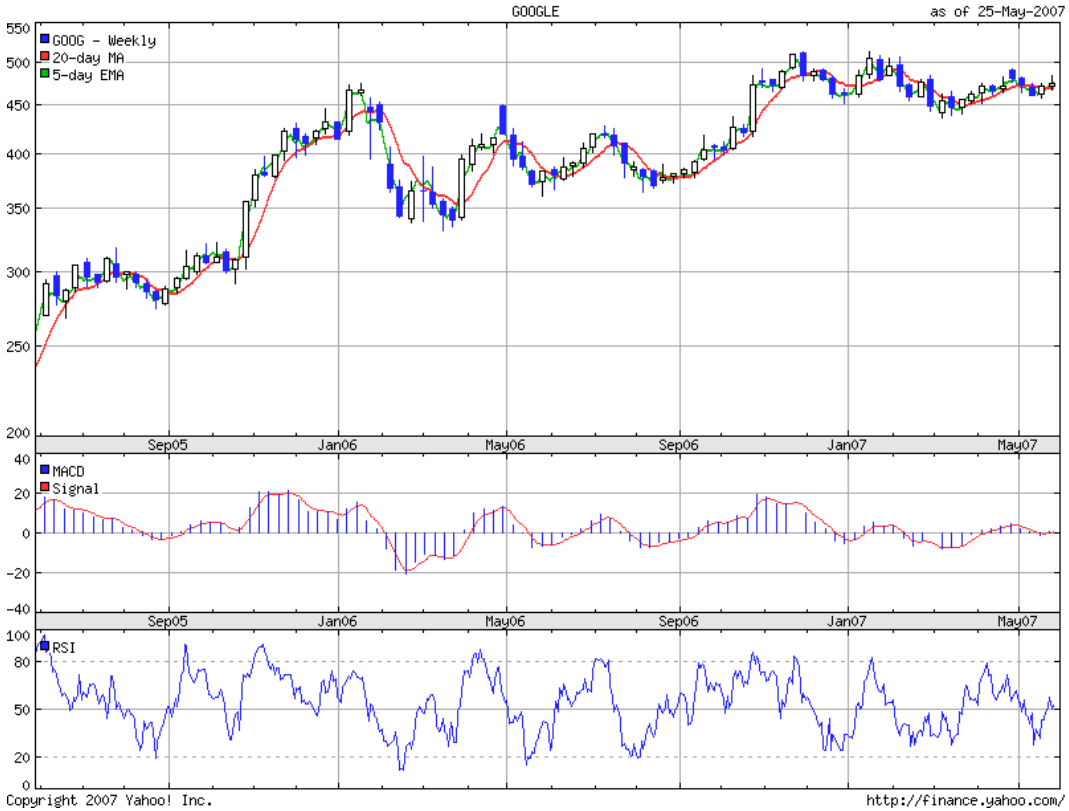
Institute of Mathematical Statistics
University of Tartu

29. June 2007

Methods of stock analysis

- Methods which predict the direction and amplitude of future movements of stock prices are divided into two groups – they are fundamental and technical analysis.
- Fundamental analysis examines the reasons of price movements – the possible value of company, growth potential, economical environment, etc.
- Technical analysis deals with the results of price movements. It searches known patterns from graphs and assumes that these patterns will recur.

Figure: Indicators of technical analysis (Google stock)



Copulas

- Copulas are functions that join one-dimensional CDF-s of random variables with their multivariate distribution.
- On the other hand, copulas are the CDF-s with marginals that have standard uniform distributions.

Definition

With given random vector (X, Y)*, its CDF* $H(x, y)$ *and marginal CDF-s* $F(x)$ *ja* $G(y)$*, we call a function* $C(u, v)$*, which is defined as* $C(u, v) = H(F^{-1}(u), G^{-1}(v))$*, a* **copula**.

Definition of some copulas

Definition

Function $C(u, v)$ which is given by $C(u, v) = B(Q^{-1}(u), Q^{-1}(v), r)$, where B is a CDF of bivariate normal distribution, Q^{-1} is an inverse of CDF of normal distribution and r is a linear correlation coefficient, is called a **Gaussian copula**.

Definition

Let the ϕ be a convex function given in $(0, \infty)$ which is increasing in $(0, 1]$ and $\phi(1) = 0$. Let the inverse of ϕ denoted as a ϕ^{-1}. Then we call a function

$$C_\phi(u, v) = \phi^{-1}(\phi(u) + \phi(v)), \text{ where } u, v \in (0, 1]$$

as an **Archimedean copula** *and a function ϕ is its generator.*

The problem of classification

- Classification is a categorization of the given object into one of given classes. Description of the object and its possible classes are given to the classifier as an input and it has to decide the correct class of the object.
- Let the description of the object be given with the feature vector $x \in \mathbb{R}^n$ and the set of possible classes be denoted as follows $\mathcal{Y} = \{0, 1, \ldots, k-1\}$. Now we can present the classifier as a function g:

$$g : \mathbb{R}^n \to \mathcal{Y}. \tag{1}$$

Marginal methods

Definition

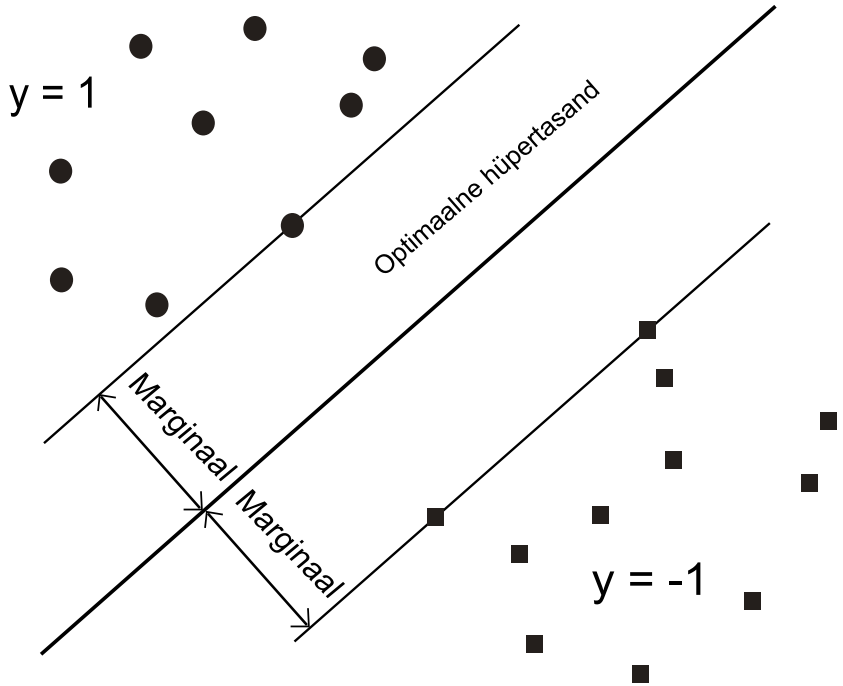
Let us have a random multivariate sample $(x_1, \ldots, x_l), x_i \in \mathbb{R}^n$ *with a given classes* $y_i \in \{-1, 1\}$ *(we know, into which class each element belongs to). If there exists a vector* ϕ *and a constant* c *such that* $y_i(\phi' x_i + c) \geq 0$ *for each pair of* (x_i, y_i) *then we say that we have a* **linearly separable sample**.

Marginal methods

- Let us have a hyperplane that is located between the two classes of sample elements and its distance from each class is maximal. Assume that this hyperplane could be the best classifier for the linearly separable sample.
- For each ϕ we define a $c_1(\phi) = \min_{y_i=1} \langle x_i, \phi \rangle$, $c_2(\phi) = \max_{y_j=-1} \langle x_j, \phi \rangle$ and a vector ϕ_0, which maximizes the equation $\rho(\phi) = \frac{c_1(\phi)-c_2(\phi)}{2}, \quad |\phi| = 1$
- Variable $\rho(\phi)$ is a distance between the hyperplane and the nearest sample element, lets call it a **marginal**.
- Such ϕ_0 and a constant $c_0 = \frac{c_1(\phi_0)+c_2(\phi_0)}{2}$ define the **optimal hyperplane** which has a maximal marginal.

Marginal methods

Figure: Optimal hyperplane divides the sample and has a maximal marginal.



Marginal methods

- we can find an optimal hyperplane by solving the dual equation of Lagrange functional.
- It comes out that only some elements from the sample define the optimal hyperplane – they are the ones that are the nearest to this hyperplane. These elements are called as a **support vectors** and they are denoted by SV.

Support Vector Machines

The idea of Support Vector Machines (SVM) is simple:

- Using some nonlinear mapping $\Phi : \mathbb{R}^n \to Z$ we project the elements of the sample $x_i, i = 1, \ldots, l, x_i \in \mathbb{R}^n$ into higher-dimensional feature space Z. Then we find
- Then we find a optimal hyperplane in Z, which separates the sample $\Phi(x_1), \ldots, \Phi(x_l)$.
- This hyperplane is nonlinear in the original space of feature vector x.

Support Vector Machines

- Using SVM-s we get that the decision rule (classifier) in original space is given by $f(x, \alpha) = sign\left(\sum_{i \in SV} y_i \alpha_i^0 K(x, x_i) + b\right)$, where α_i^0 and b are parameters found by solving optimization problem and $K(x, x_i)$ is a predefined function. Function K is called **kernel** and it defines an inner product for space Z.
- The shape of kernel K is given by statistician and it depends on the type of data (it has to be positively semi-defined).
- I used Gaussian kernel: $K(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right)$ and polynomial kernel: $K(x, y) = (x'y + R)^p$.

Modeling

- we look the history of IBM corp. stock data – 6210 days which are devided into training sample (to solve the optimization problem and find the optimal hyperplane) and test sample (to test the accuracy of the model.
- for the model we take the volume and price of the stock data in days $i, i-1, \ldots, i-19$ and use them as an input data. For each day we have:
  - *ln* ratio of closing prices in sequential days;
  - *ln* ratio of opening and closing price in one day;
  - *ln* ratio of maximal price and closing price in one day;
  - *ln* ratio of minimal price and closing price in one day;
  - *ln* ratio of volumes in sequential days.
- Trading day i belongs to class $y_i = 1$, if the stock price in next day increased and into class $y_i = -1$, if the price decreased.

Classification

- We used Gaussian and Archimedean copula and SVM-s with Gaussian and polynomial kernel for data classification.
- The dimension of the input vector was different: for SVM-s it was 100, for copulas 2 and 3.
- We trained the models using training sample and tested them on test sample, but non of models gave good results. For all of them the accuracy of classifying test data was below 54%.