Computational Intelligence: Methods and Applications

Lecture 23
Logistic discrimination
and support vectors

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Logistic discrimination

Basic assumption of the logistic model: logarithm of the ratio of class distribution is a linear function:

$$\log\left(\frac{P(\mathbf{X}|\omega_1)}{P(\mathbf{X}|\omega_2)}\right) = \mathbf{W}^T\mathbf{X} + W_0$$

This is exact when class distributions are normal (Gaussian) with equal covariance matrices, and for some discrete data distributions.
Since these probabilities sum to 1, using the Bayesian formula $P(\omega|X) = P(X|\omega)\, P(\omega)/P(X)$, the model is equivalent to:

$$P(\omega_2|\mathbf{X}) = \frac{1}{1+\exp\left(\mathbf{W}^T\mathbf{X}+W_0'\right)}; \quad W_0' = W_0 + \log\frac{P(\omega_1)}{P(\omega_2)}$$

$$P(\omega_1|\mathbf{X}) = \frac{\exp\left(\mathbf{W}^T\mathbf{X}+W_0'\right)}{1+\exp\left(\mathbf{W}^T\mathbf{X}+W_0'\right)} = 1 - P(\omega_2|\mathbf{X})$$

Logistic DA

Classification rule is therefore:

$$\Lambda = \frac{P(\omega_1|\mathbf{X})}{P(\omega_2|\mathbf{X})} > 1 \text{ Then Class } \omega_1 \text{ Else } \omega_2$$

$$\text{or } \mathbf{W}^T\mathbf{X} + W_0' > 0 \text{ Then Class } \omega_1 \text{ Else } \omega_2$$

This time probabilities (observations) are non-linear functions of parameters W; usually iterative procedures based on maximization of likelihood of generation of the observed data are used, equivalent to:

$$L(\mathbf{W}, W_0) = \prod_{\mathbf{X}\in\omega_1} P(\omega_1|\mathbf{X}) \prod_{\mathbf{X}\in\omega_2} P(\omega_2|\mathbf{X})$$

Using logistic functions for $P(\omega|X)$ and calculating gradients in respect to W leads to a non-linear optimization problem.

This is implemented in WEKA/YALE, giving usually better results than LDA at some increase computational costs.

WEKA Logistic voting

Similar results to LDA

Whole data:
=== Confusion Matrix ===

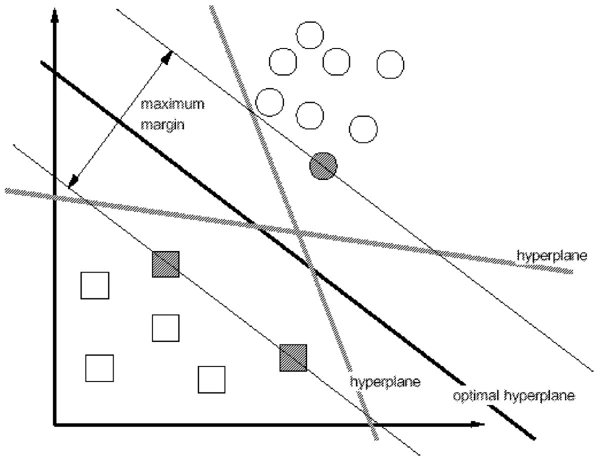
a	b	<= classified as
260	7	a = democrat
5	163	b = republican

10xCV results

a	b	<= classified as
258	9	a = democrat
9	159	b = republican

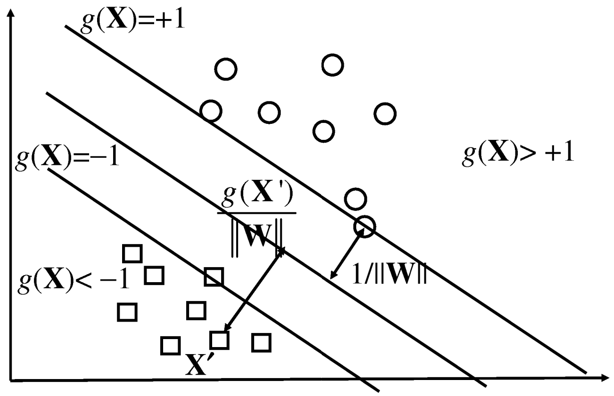
Decision trees give better results in this case, perhaps one hyperplane is not sufficient.

Maximization of margin 1



Among all discriminating hyperplanes there is one that is clearly better.

Maximization of margin 2



$g(\mathbf{X})=\mathbf{W}^T\mathbf{X}+W_0$ is the discriminant function, $g(\mathbf{X})/||\mathbf{W}||$ is the distance.
The best discriminating hyperplane should maximize the distance between the $g(\mathbf{X})=0$ plane and the data samples that are near to it.

Maximization of margin 3

Maximize the distance $g_W(\mathbf{X})/||\mathbf{W}||$ between the plane $\mathbf{W}$ and data samples, or maximize the value of discriminant $g_W(\mathbf{X})$ for $||\mathbf{W}||=1$

Find vectors $\mathbf{X}^{(i)}$ that are close to $\mathbf{W}$ hyperplane in *d* dimensions:

$$\mathbf{X}^{(i)} = \arg\min_{\mathbf{X}} g_W(\mathbf{X}) = \min_{\mathbf{X}}\left(\mathbf{W}^T\mathbf{X}+W_0\right)$$

For these vectors find $\mathbf{W}$ giving maximum distance

$$\max_{\mathbf{W}} D\left(\mathbf{W},\mathbf{X}^{(i)}\right) = \max_{\mathbf{W}} g_W\left(\mathbf{X}^{(i)}\right)\Big/\|\mathbf{W}\|$$

Which vectors to choose as "support" for such calculation? Let the target values for classification be $Y(\omega_1)=+1$ and $Y(\omega_2)=-1$ and the margin *b* be the distance between W and these support vectors:

$$Y^{(i)}\frac{g_W\left(\mathbf{X}^{(i)}\right)}{\|\mathbf{W}\|} \geq b, \quad i=1..n$$

This should be true for all vectors, in a separable case.

Formulation of the problem

Setting $b||\mathbf{W}||=1$ (particular choice of b) separation conditions are:

$$Y^{(i)} g_W\left(\mathbf{X}^{(i)}\right) \geq 1, \quad i=1..n$$

These conditions define two canonical hyperplanes:

$$H_1: g_W\left(\mathbf{X}\right) = \mathbf{W}^T\mathbf{X}+W_0 = +1$$
$$H_2: g_W\left(\mathbf{X}\right) = \mathbf{W}^T\mathbf{X}+W_0 = -1$$

Distance of H_i from the H_0 separating plane $g_W(\mathbf{X})=0$ is $D(H_0,H_i)=1/||\mathbf{W}||$

Largest margin is obtained from minimization of $||\mathbf{W}||$ with $g_W(\mathbf{X})$, fulfilling the separation conditions.
This leads to a constrained minimization problem.

Minimize $||\mathbf{W}||$ with constraints $Y^{(i)} g_W\left(\mathbf{X}^{(i)}\right) \geq 1, \quad i=1..n$

Support vectors are vectors that are the closest to the separating hyperplane, most difficult to separate and most informative.

Scalar product form

In the d-dimensional space if $n > d$ the weight vector may be expresses as the combination of:

$$\mathbf{W} = \sum_{i=1}^{n} \alpha_i \mathbf{X}^{(i)}$$

It should be enough to take only d independent training vectors, so most α_i=0. Therefore the discriminant function:

$$g_{\mathbf{W}}(\mathbf{X}) = \mathbf{W}^{\mathrm{T}}\mathbf{X} = \sum_i \alpha_i \mathbf{X}^{(i)\mathrm{T}} \cdot \mathbf{X}$$

$$g_{\mathbf{W}}\left(\mathbf{X}^{(j)}\right) = \mathbf{W}^{\mathrm{T}}\mathbf{X}^{(j)} = \sum_i \alpha_i \mathbf{X}^{(i)\mathrm{T}} \cdot \mathbf{X}^{(j)}$$

$$= \sum_i \alpha_i K\left(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}\right) = \sum_i \alpha_i K_{ij}$$

The kernel matrix K_{ij} will play an important role soon ...

Lagrange form and SV

Lagrange multiplier method is used to convert constraint minimization problems into a simpler optimization problem (here X includes X_0=1):

$$L(\mathbf{W}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{W}\|^2 - \sum_{i=1}^{n} \alpha_i \left[Y^{(i)} g_{\mathbf{W}}\left(\mathbf{X}^{(i)}\right) - 1 \right], \alpha_i \geq 0, \quad i = 1..n$$

where $\boldsymbol{\alpha}$ are Lagrangian multipliers - free positive parameters, and summation runs over the number of all training samples *n*.

Minimization of the Lagrangian function over W increases margin.

Suppose that $X^{(i)}$ is misclassification, then the second term $g(\mathbf{X}^{(i)})-1$ in the Lagrangian is negative, and large α_i will create a large contribution to $L(\mathbf{W},\boldsymbol{\alpha})$; this will be decreased by changing W to remove the error. Therefore ||W|| should be minimized and $\boldsymbol{\alpha}$ maximized, but only for vectors for which $g(\mathbf{X}^{(i)})-1=0$, called Support Vectors (SV).

This leads to the search for the saddle point, not minima; to simplify it W parameters are replaced by $\boldsymbol{\alpha}$.

Scalar product discriminant

Differentiating in respect to W and W_0 gives:

$$\frac{\partial L(\mathbf{W}, \boldsymbol{\alpha})}{\partial W_0} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i Y^{(i)} = 0$$

$$\frac{\partial L(\mathbf{W}, \boldsymbol{\alpha})}{\partial \mathbf{W}} = 0 \Rightarrow \mathbf{W} = \sum_{i=1}^{n} \alpha_i Y^{(i)} \mathbf{X}^{(i)}$$

Interesting! W is now a linear combination of input vectors!

Makes sense, since a component W_Z of $W=W_Z+W_X$ that does not belong to the space spanned by $\mathbf{X}^{(i)}$ vectors has no influence on the discrimination process, because $\mathbf{W}_Z^{\mathrm{T}}\mathbf{X}=0$.

Inserting $\mathbf{W}$ in the discriminant function:

$$g(\mathbf{X}) = \mathbf{W}^{\mathrm{T}} \cdot \mathbf{X} + W_0 = \sum_{i=1}^{n} \alpha_i Y^{(i)} \mathbf{X}^{(i)\mathrm{T}} \cdot \mathbf{X} + W_0$$

for support vector $Y^{(i)}g(\mathbf{X}^{(i)})=1$, so $W_0 = Y^{(i)} - \mathbf{W}^{\mathrm{T}} \cdot \mathbf{X}^{(i)}$

Lagrangian in dual form

Substituting W into the Lagrangian leads to a maximization of a dual form (**X** here may be *d*+1 dim or *d*-dim, it does not matter):

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \alpha_i Y^{(i)} \sum_{j=1}^{n} \alpha_j Y^{(j)} \mathbf{X}^{(i)\mathrm{T}} \cdot \mathbf{X}^{(j)}$$

$$\sum_{i=1}^{n} \alpha_i Y^{(i)} = 0; \quad \alpha_i \geq 0; \quad i = 1..n$$

In this form optimization criterion is expressed as inner products of support vectors, and is now **maximized** subject to constraints.

Initially number of parameters is equal to the number of patterns *n*, usually much bigger than dimensionality *d*, but the final number of non-zero $\boldsymbol{\alpha}$ may be small.

This type of quadratic minimization problem has a unique solution!

Popular approach: SMO, Sequential Minimal Optimization algorithm for Quadratic Programming, fast and accurate.