PA196: Pattern Recognition

03. Linear discriminants

Dr. Vlad Popovici
popovici@iba.muni.cz

Institute of Biostatistics and Analyses
Masaryk University, Brno

Outline

Outline

1. Linear Discriminant Analysis (cont'd)
   - LDA, QDA, RDA
   - LD subspace
   - LDA: wrap-up
2. Logistic regression
3. Large margin (linear) classifiers
   - Linearly separable case
   - Soft margins (Non-linearly separable case)

LDA

Remember (first lecture):

- Bayes decision: assign $\mathbf{x}$ to the class with maximum a posteriori probability
- let there be K classes denoted $g_1, \ldots, g_K$, with corresponding priors $P(g_i)$
- the posteriors are:

$$P(g_i|\mathbf{x}) = \frac{p(\mathbf{x}|g_i)P(g_i)}{\sum_i^K p(x|g_i)P(g_i)} \propto p(\mathbf{x}|g_i)P(g_i)$$

- decision function (for class g_i vs class g_j) arise from log odds-ratios (for example):

$$\log \frac{P(g_i|\mathbf{x})}{P(g_j|\mathbf{x})} = \log \frac{p(\mathbf{x}|g_i)}{p(\mathbf{x}|g_j)} + \frac{P(g_i)}{P(g_j)} \begin{cases} > 0, & \text{predict } g_i \\ < 0, & \text{predict } g_j \end{cases}$$

Under the *assumption* of Gaussian class-conditional densities:

$$p(\mathbf{x}|g) = \frac{1}{(2\pi)^d |\Sigma_g|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)\right]$$

($|\Sigma|$ is the determinant of covariance matrix Σ) the decision function becomes

$$h_{ij}(\mathbf{x}) = \log \frac{P(g_i|\mathbf{x})}{P(g_j|\mathbf{x})} = (\mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}) - (\mathbf{x}^t \mathbf{W}_j \mathbf{x} + \mathbf{w}_j^t \mathbf{x} + w_{j0})$$

where

$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}, \quad \mathbf{w}_i = \Sigma_i^{-1}\mu_i$$

and

$$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1} \mu_i - \frac{1}{2}\log|\Sigma_i| + \log P(g_i)$$

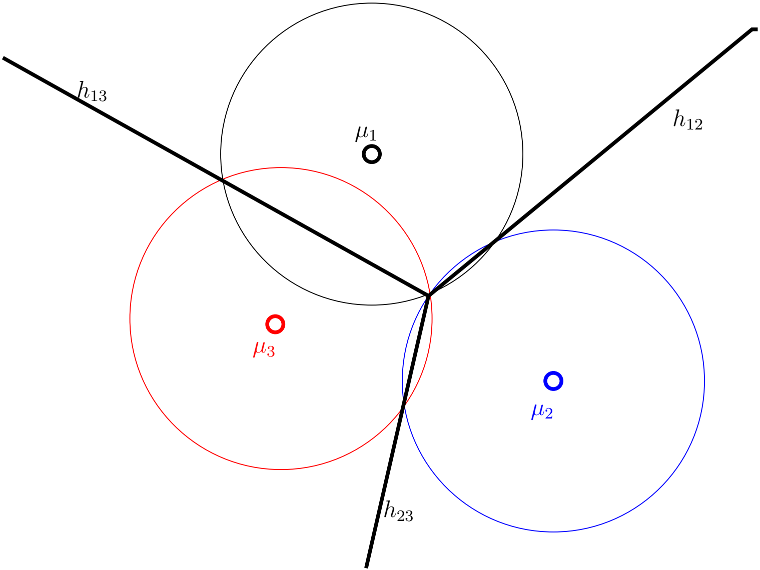
Simplest LDA

If $\Sigma_i = \Sigma_j = \sigma^2\mathbf{I}$ ("spherical" covariance matrices)

$$h_{ij}(\mathbf{x}) = \mathbf{w}_{ij}(\mathbf{x} - \mathbf{x}_0)$$

where

$$\mathbf{w}_{ij} = \mu_i - \mu_j, \quad \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \log \frac{P(g_i)}{P(g_j)}(\mu_i - \mu_j)$$


h_{13}
h_{12}
μ_1
μ_3
μ_2
h_{23}

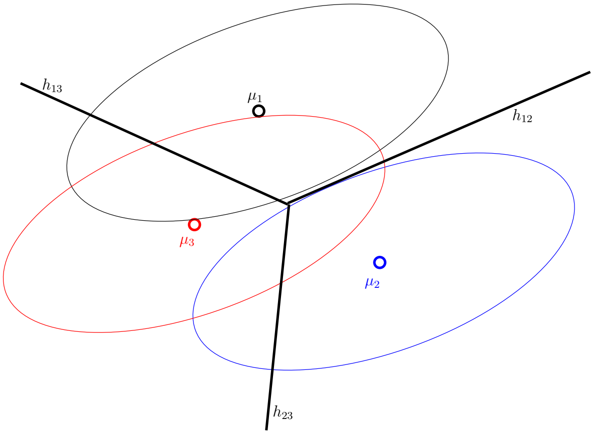
Classical LDA

If all classes share a common covariance matrix, $\Sigma_i = \Sigma$, the decision function becomes

$$h_{ij}(\mathbf{x}) = \mathbf{w}^t(\mathbf{x} - \mathbf{x}_0)$$

where

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j),\ \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{1}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P(g_i)}{P(g_j)}(\mu_i - \mu_j)$$


h_{13}
μ_1
h_{12}
μ_3
μ_2
h_{23}

Estimation of LDA parameters

- we are given $\{(\mathbf{x}_i, g_i), i = 1, \ldots, n\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $g_i \in \{g_1, \ldots, g_K\}$.
- priors: $\hat{P}(g_i) = n_i/n$ where n_i is the number of elements of class g_i in the training set
- mean vectors: $\hat{\mu}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in g_i} \mathbf{x}$
- covariance matrix: $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in g_k} (\mathbf{x} - \hat{\mu}_k)(\mathbf{x} - \hat{\mu}_k)^t / (n - K)$

Quadratic Discriminant Analysis

Class-conditional probabilities are general Gaussians and the decision function has the form:

$$h_{ij}(\mathbf{x}) = \log \frac{P(g_i|\mathbf{x})}{P(g_j|\mathbf{x})} = (\mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}) - (\mathbf{x}^t \mathbf{W}_j \mathbf{x} + \mathbf{w}_j^t \mathbf{x} + w_{j0})$$

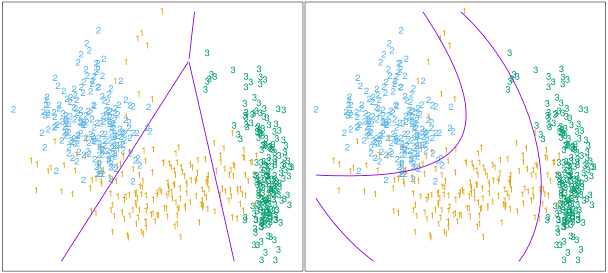
where

$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}, \quad \mathbf{w}_i = \Sigma_i^{-1}\mu_i$$

and

$$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\log|\Sigma_i| + \log P(g_i)$$

LDA and QDA



Hastie et al: The Elements of Statistical Learning - chpt 4

Note: a similar boundary to QDA could be obtained by applying LDA in an augmented space with axes $x_1, \ldots, x_d$, $x_1 x_2, \ldots, x_{d-1} x_d$, $x_1^2, \ldots, x_d^2$

Regularized DA: between LDA and QDA

Combine the pooled covariance with class-specific covariance matrices, and allow the pooled covariance the be *more spherical* or *more general*:

$$\hat{\Sigma}_k(\alpha, \gamma) = \alpha \hat{\Sigma}_k + (1 - \alpha) \left[\gamma \hat{\Sigma} + (1 - \gamma) \hat{\sigma}^2 \mathbf{I} \right]$$

- $\alpha = 1$: QDA; $\alpha = 0$: LDA
- $\gamma = 1$: general covariance matrix; $\gamma = 0$: spherical covariance matrix
- α and γ must be optimized

Implementation of LDA

- use diagonalization of the covariance matrices (either pooled or class-specific), which are *symmetric and positive definite*:

$$\Sigma_i = \mathbf{U}_i \mathbf{D}_i \mathbf{U}_i^t$$

  where $\mathbf{U}_i$ is a $d \times d$ orthonormal matrix and D_i is a diagonal matrix with eigenvalues $d_{ik} > 0$ on the diagonal
- the ingredients for the decision functions become:

$$(\mathbf{x} - \mu_i)^t \Sigma_i^{-1} (\mathbf{x} - \mu_i) = [\mathbf{U}_i^t (\mathbf{x} - \mu_i)]^t \mathbf{D}_i^{-1} [\mathbf{U}_i^t (\mathbf{x} - \mu_i)]$$

  and

$$\log |\Sigma_i| = \sum_k \log d_{ik}$$

Implementation of LDA, cont'd

A possible 2-step procedure for LDA classification (common covariance matrix $\Sigma = \mathbf{UDU}^t$):

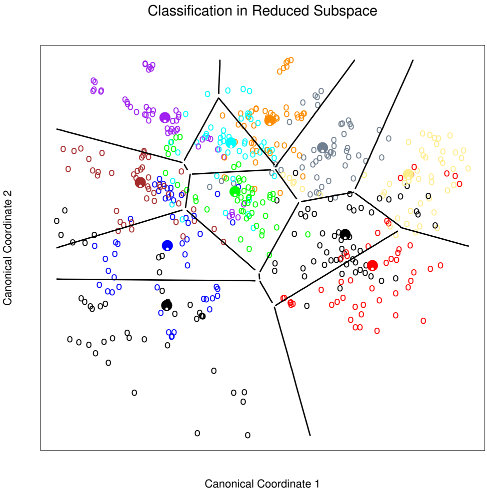
1. "sphere" the data: $\mathbf{X}^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^t\mathbf{X}$
2. assign a sample $\mathbf{x}$ to the *closest centroid* in transformed space, modulo the effect of the priors

Outline

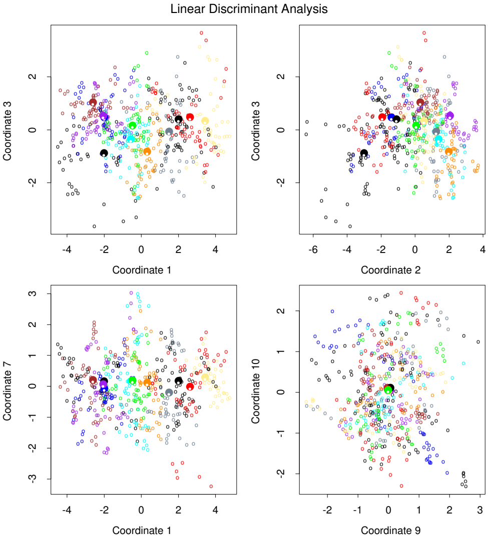
- the centroids μ_i $i = 1, \ldots, K$ lie in an affine subspace of dimension at most $K - 1 < d$
- any dimension orthogonal to this subspace does not influence the classification
- the classification is carried out in a low dimensional space, hence we have a dimensionality reduction
- the subspace axes can be found sequentially, using Fisher's criterion (find directions that maximally separate the centroids with respect to the variance)
- this is essentially the same as *Principal Component Analysis*

1. compute $\mathbf{M}$ the $K \times d$ matrix of class centroids (by rows) and the common covariance matrix $\mathbf{W}$ (within-class covariance)
2. compute $\mathbf{M}^* = \mathbf{M}\mathbf{W}^{-\frac{1}{2}}$ (using eigen-decomposition of $\mathbf{W}$)
3. compute $\mathbf{B}^*$ – the covariance matrix of $\mathbf{M}^*$ (between-class covariance matrix), and its eigen-decomposition $\mathbf{B}^* = \mathbf{V}^*\mathbf{D}_B\mathbf{V}^{*t}$
4. the columns of $\mathbf{V}^*$ (ordered from largest to smallest eigen-value d_{Bi}) give the coordinates of the optimal subspaces

- the i–th *discriminant variable (canonical variable)* is given by $Z_i = (\mathbf{W}^{-\frac{1}{2}} v_i^*)^t\mathbf{X}$



Hastie et al. - The Elements of Statistical Learning - chpt. 4



Hastie et al. - The Elements of Statistical Learning - chpt. 4

Outline

- LDA, FDA and MSE regression with a particular coding of class labels, lead to equivalent solutions (separating hyperplane)
- LDA (QDA) is the optimal classifier in the case of Gaussian class-conditional distributions
- LDA can be used to project data into a lower dimensional space for visualization
- LDA derivation assumes Gaussian densities, but FDA does not
- LDA is naturally extended to multiple classes

Outline

1. Linear Discriminant Analysis (cont'd)
   - LDA, QDA, RDA
   - LD subspace
   - LDA: wrap-up

2. Logistic regression

3. Large margin (linear) classifiers
   - Linearly separable case
   - Soft margins (Non-linearly separable case)

Idea: model the posterior probabilities as linear functions in $\mathbf{x}$ and ensure they sum up to 1.

For K classes $g_1, \ldots, g_K$:

$$\log \frac{P(g_i|\mathbf{x})}{P(g_K|\mathbf{x})} = \langle \mathbf{w}_i, \mathbf{x} \rangle + w_{i0}, \qquad \forall i = 1, \ldots, K-1$$

which leads to

$$P(g_i|\mathbf{x}) = \frac{\exp(\langle \mathbf{w}_i, \mathbf{x} \rangle + w_{i0})}{1 + \sum_{j=1}^{K-1} \exp(\langle \mathbf{w}_j, \mathbf{x} \rangle + w_{j0})}, \quad i = 1, \ldots, K-1$$

$$P(g_K|\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\langle \mathbf{w}_j, \mathbf{x} \rangle + w_{j0})}$$

- the transformation $p \mapsto \log[p/(1-p)]$ is called *logit transform*
- the choice of reference class (K in our case) is purely a convention
- the set of parameters of the model: $\theta = \{\mathbf{w}_1, w_{10}, \dots, \mathbf{w}_{K-1}, w_{K-1,0}\}$
- the *log-likelihood* is

$$L(\theta) = \sum_{i=1}^{n} \log P(g_i | x_i; \theta)$$

Binary case ($K = 2$):

- take the classes to be encoded in response variables y_i: $y_i = 0$ for class g_1 and $y_i = 1$ for class g_2.
- a single posterior probability is needed: let it be

$$P(y = 0|\mathbf{x}) = \frac{\exp(\langle \mathbf{w}, \mathbf{x} \rangle + w_0)}{1 + \exp(\langle \mathbf{w}, \mathbf{x} \rangle + w_0)}$$

- the likelihood function becomes:

$$L(\theta = \{\mathbf{w}, w_0\}) = \sum_{i=1}^{n} [y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) - \log(1 + \exp(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0)]$$

- using $\mathbf{z} = [1, \mathbf{x}]$ and $\mathbf{a} = [w_0, \mathbf{w}]$,

$$L(\mathbf{a}) = \sum_{i=1}^{n} [y_i \langle \mathbf{a}, \mathbf{z}_i \rangle - \log(1 + \exp(\langle \mathbf{a}, \mathbf{z}_i \rangle))]$$

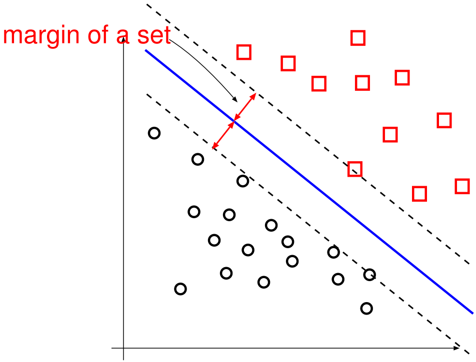
- objective: find $\mathbf{a}* = \arg\max_{\mathbf{a}} L(\mathbf{a})$
- $\frac{\partial L(\mathbf{a})}{\partial \mathbf{a}} = \sum_{i=1}^{n} \mathbf{z}_i (y_i - P(y_i = 0 | \mathbf{z}_i))$
- at a (local) extremum, $\frac{\partial L(\mathbf{a})}{\partial \mathbf{a}} = 0$ which is the system of equations to be solved for $\mathbf{a}$
- the solution can be found by a Newton-Raphson procedure (iteratively reweighted least squares)

A few remarks on logistic regression:

- brings the tools from linear regression to pattern recognition problems
- can be used to identify those input variables that *explain* the output
- its predictions can be interpreted as posterior probabilities
- by introducing a penalty term, variable selection can be embedded into the model construction - we'll see it later!
- both LDA and logistic regression use a linear form for the log-posterior odds ($\log(P(g_i|x)/P(g_K|x))$); LDA assumes the posteriors to be Gaussians, while logistic regression assumes they only lead to linear log-posterior odds

Outline

- there are theoretical considerations to justify the goal of maximizing the margin achieved by the separating hyperplane
- intuitively, larger the margin, more "room" for noise in the data and hence, better generalization
- let a training set be $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ with $y_i = \pm 1$
- the margin of a point $\mathbf{x}_i$ with respect to the boundary function h is $\gamma_i = y_i h(\mathbf{x}_i)$
- it can be shown that the maximal error attained by h is upper bounded by a function of $\min(\gamma_i)$ (however, the bound might not be tight)


margin of a set

Outline

1. Linear Discriminant Analysis (cont'd)
   - LDA, QDA, RDA
   - LD subspace
   - LDA: wrap-up
2. Logistic regression
3. Large margin (linear) classifiers
   - Linearly separable case
   - Soft margins (Non-linearly separable case)

- consider the dataset $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be linearly separable, i.e. $\gamma_i > 0$
- we will consider linear classifiers $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + w_0$ (with the predicted class $\text{sign}(h(\mathbf{x}))$)
- if the (functional) margin achieved is 1, then $\gamma_i \geq 1$
- then, the geometric margin is the normalized functional margin: $1/||\mathbf{w}||$, hence:

Proposition

The hyperplane $(\mathbf{w}, w)$ that solves the optimization problem

$$\begin{aligned} \text{minimize}_{\mathbf{w}, w_0} \quad & \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle, \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1, \; i = 1, \ldots, n \end{aligned}$$

realizes the maximal margin hyperplane with geometric margin $\gamma = 1/||\mathbf{w}||$.

Solving the constrained optimization:

- let the objective function be $f(\mathbf{w})$ and the equality constraints $h_i(\mathbf{w}) = 0$ for $i = 1, \ldots, m$, then the *Lagrangian function* is

$$L(\mathbf{w}, \beta) = f(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w})$$

- a necessary and sufficient condition for $\mathbf{w}^*$ to be a solution of the optimization problem (f continuous and convex, h_i continuous and differentiable) is

$$\frac{\partial L(\mathbf{w}^*, \beta^*)}{\partial \mathbf{w}} = 0$$
$$\frac{\partial L(\mathbf{w}^*, \beta^*)}{\partial \beta} = 0$$

for some values of β^*

For a constrained optimization with a domain $\Omega \subseteq \mathbb{R}^n$:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{w}), \quad \mathbf{w} \in \Omega \\ \text{subject to} \quad & g_i(\mathbf{w}) \leq 0,\ i = 1, \ldots, k \\ & h_i(\mathbf{w}) = 0\ i = 1, \ldots, m \end{aligned}$$

the Lagrangian function has the form

$$L(\mathbf{w}, \alpha, \beta) = f(\mathbf{w}) + \sum_i \alpha_i g_i(\mathbf{w}) + \sum_j \beta_j h_j(\mathbf{w})$$

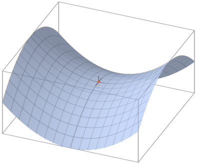
with α_i and β_j being the Lagrange multipliers.

Karush-Kuhn-Tucker (KKT) optimality conditions for a convex optimization problem: for a solution $\mathbf{w}^*$ and corresponding multipliers α^* and β^*,

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= 0 \\ \frac{\partial L}{\partial \beta} &= 0 \\ \alpha_i^* g_i(\mathbf{w}^*) &= 0 \\ g_i(\mathbf{w}^*) &\leq 0 \\ \alpha_i^* &\geq 0 \end{aligned}$$

- for active constraints ($g_i(\mathbf{w}) = 0$), $\alpha_i > 0$; for inactive constraints ($g_i(\mathbf{w}) < 0$), $\alpha_i = 0$
- α_i can be seen as the sensitivity of f to the active constraint

Duality of convex optimization:



- the solution is a *saddle point*
- **w** are the *primal* variables
- Lagrange multipliers are the *dual* variables
- solving the dual optimization may be simpler: the Lagrange multipliers are the main variables, so set to 0 the derivatives wrt to **w** and substitute the result into the Lagrangian
- the resulting function contains only dual variables and must be *maximized* under simpler constraints

...and back to our initial problem:

- the primal Lagrangian is

$$L(\mathbf{w}, w_0, \alpha) = \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle - \sum_{i=1}^{n} \alpha_i \left[y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) - 1 \right]$$

- from KKT conditions, $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$ and $\sum_{i=1}^{n} y_i \alpha_i = 0$
- which leads to the dual Lagrangian

$$L(\mathbf{w}, w_0, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

Proposition

If α^* is the solution of the quadratic problem

$$\text{maximize} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \; i = 1, \ldots, n$$

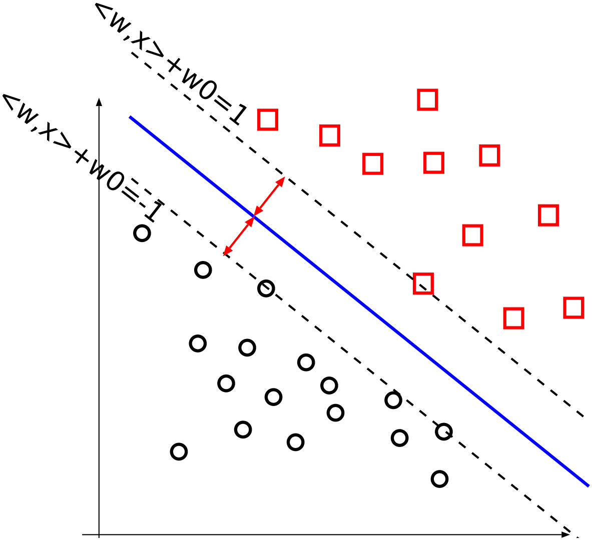
then the vector $\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i^* \mathbf{x}_i$ realizes the maximal margin hyperplane with the geometric mean $1/||\mathbf{w}^*||$.

- in the dual formulation, w_0^* still needs to be specified, so

$$w_0^* = -\frac{1}{2}\left(\max_{\gamma_i=-1} \{\langle \mathbf{w}^*, \mathbf{x}_i \rangle\} + \min_{\gamma=1} \{\langle \mathbf{w}^*, \mathbf{x}_i \rangle\} \right)$$

- from the KKT conditions: $\alpha_i^*[y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + w_0^*) - 1] = 0$, so only for $\mathbf{x}_i$ lying on the margin, the $\alpha_i^* \neq 0$
- those $\mathbf{x}_i$ for which $\alpha_i^* \neq 0$ are called support vectors
- the optimal hyperplane is a linear combination of support vectors:

$$h(\mathbf{x}) = \sum_{i \in SV} y_i \alpha_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle + w_0^*$$


<w,x>+w0=1
<w,x>+w0=-1

- the margin achieved is

$$\gamma = \frac{1}{\|\mathbf{w}^*\|} = \left(\sum_{i \in SV} \alpha_i^* \right)^{-\frac{1}{2}}$$

- (a leave-one-out) estimate of the generalisation error is the proportion of support vectors of the total training sample size

$$\frac{\sharp SV}{n}$$

Outline

2-Norm soft margin

- introduce the *slack variables* ξ and allow "softer" margins:

$$\begin{aligned} \text{minimize}_{\mathbf{w},w_0,\xi} \quad & \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + C\sum_{i=1}^{n}\xi^2, \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + w_0) \geq 1 - \xi,\ i = 1, \dots, n \\ & \xi_i \geq 0,\ i = 1, \dots, n \end{aligned}$$

  for some $C > 0$
- theory suggests optimal choice for C: $1/\max_i\{\|\mathbf{x}_i\|^2\}$, but in practice C is selected by testing various values
- the problem is solved in dual space and the margin achieved is $\left(\sum_{i\in SV}\alpha_i^* - \|\alpha^*\|^2/C\right)^{-1/2}$

1-Norm soft margin

- optimization problem:

$$\begin{aligned} \text{minimize}_{\mathbf{w}, w_0, \xi} \quad & \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^{n} \xi, \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geq 1 - \xi, \; i = 1, \dots, n \\ & \xi_i \geq 0, \; i = 1, \dots, n \end{aligned}$$

  for some $C > 0$
- this results in "box constraints" on α_i: $0 \leq \alpha_i \leq C$
- non-zero slack variables correspond to $\alpha_i = C$ and to points with geometric margin less than $1/||\mathbf{w}||$

Wrap-up

- LDA and MSE-based methods lead to similar solutions, even though they are derived under different assumptions
- LDA (and FDA) assign the vectors $\mathbf{x}$ to the closest centroid, in a transformed space
- logistic regression and LDA model the likelihood as a linear function
- the predicted values from logistic regression can be interpreted as posterior probabilities
- margin optimization provides an alternative approach