# Overview for Week 5

THE AUSTRALIAN NATIONAL UNIVERSITY

**Support Vector Classification**

Hard Margin, Optimization Problem, Dual Objective Function, Soft Margin, Kernel Formulation

**Support Vector Regression**

$\varepsilon$-insensitive loss, Optimization Problem, Dual Objective Function, Soft Margin, Kernel Formulation

**Novelty Detection**

Basic Idea, Optimization Problem, Applications

**$\nu$-Trick**

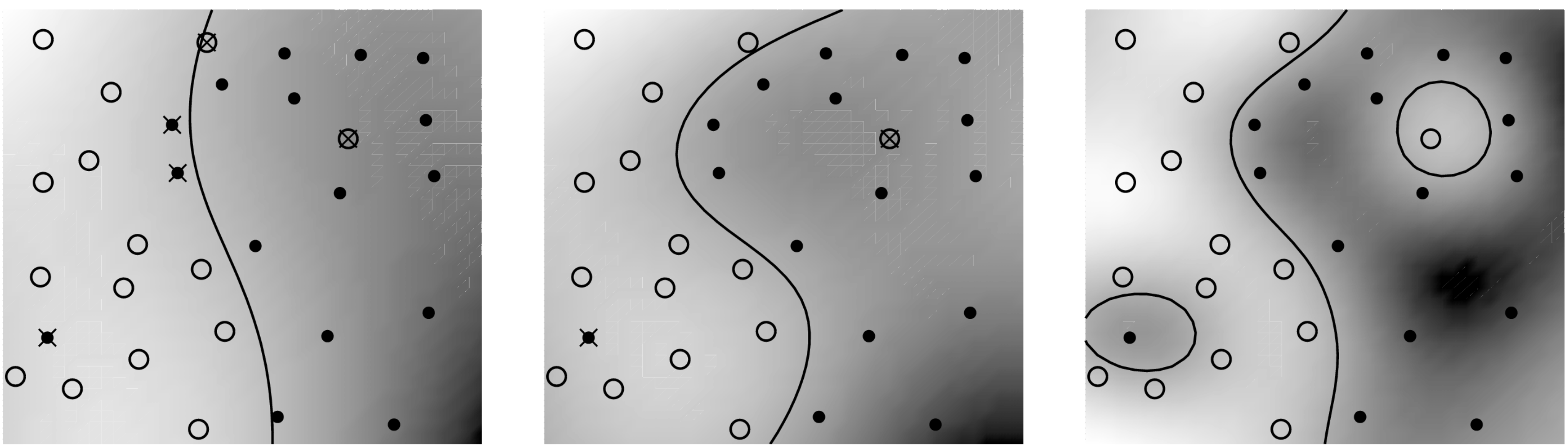
How to adjust the number of training errors automatically, optimization problems, rules of thumb

# Classification

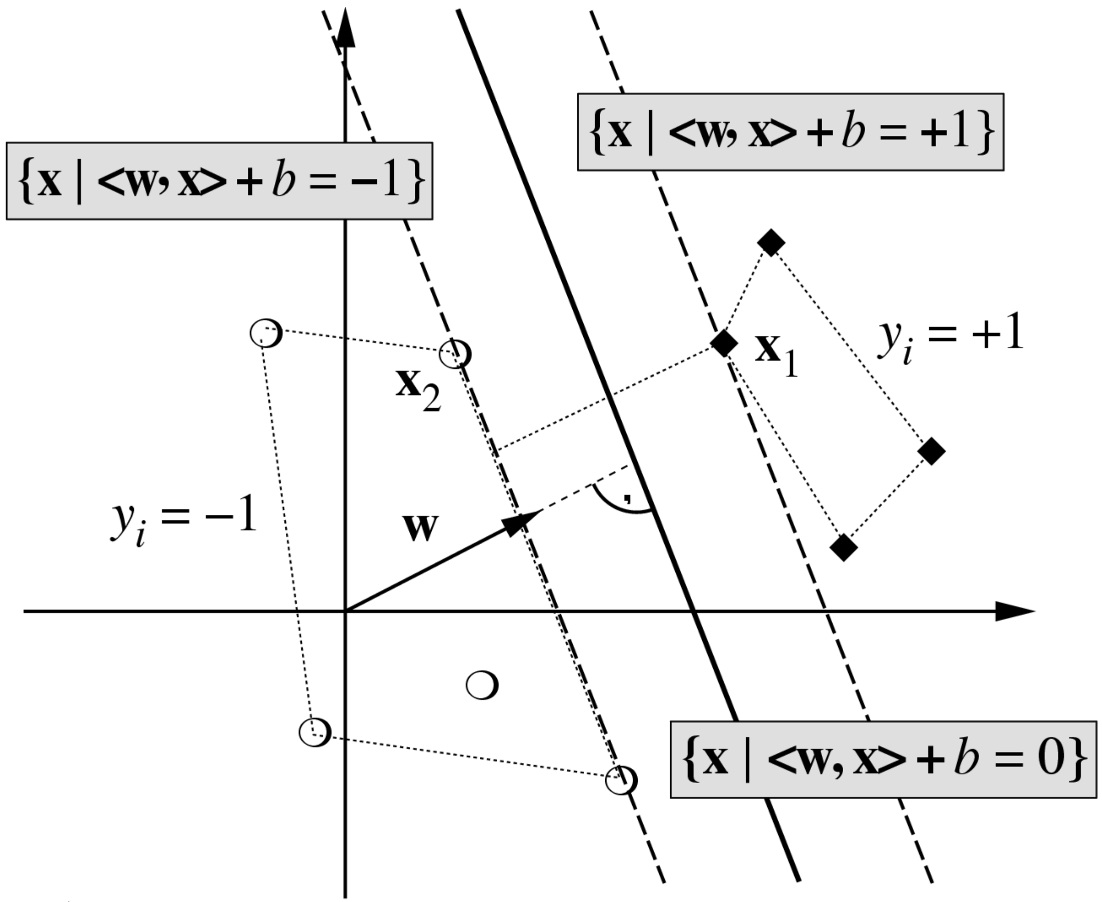
**Data:** Pairs of observations $(\mathbf{x}_i, y_i)$ generated from some distribution $\mathrm{P}(\mathbf{x}, y)$, e.g., (blood status, cancer), (credit transaction information, fraud), (sound profile of jet engine, defect)

**Task: Predict $y$ given $\mathbf{x}$** at a new location.

Modification: find a function $f(\mathbf{x})$ that does the task.



# Optimal Separating Hyperplane




Note:

$$\langle w, x_1 \rangle + b = +1$$
$$\langle w, x_2 \rangle + b = -1$$
$$\Rightarrow \quad \langle w, (x_1 - x_2) \rangle = 2$$
$$\Rightarrow \quad \left\langle \frac{w}{\|w\|}, (x_1 - x_2) \right\rangle = \frac{2}{\|w\|}$$

Minimize $\frac{1}{2}\|w\|^2$ subject to $y_i(\langle w, x_i \rangle + b) \geq 1$ for all $i$.

# Optimization Problem

THE AUSTRALIAN NATIONAL UNIVERSITY

## Linear Function

$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$

## Classification Constraint

To ensure that all $f(\mathbf{x}_i)$ lie on the "right" side of the margin we require that

$$\begin{aligned} f(\mathbf{x}_i) &= \langle \mathbf{w}, \mathbf{x}_i \rangle + b > 1 && \text{for } y_i = 1 \\ f(\mathbf{x}_i) &= \langle \mathbf{w}, \mathbf{x}_i \rangle + b < -1 && \text{for } y_i = -1 \end{aligned}$$

## Maximum Margin

For maximum margin we want to minimize $\frac{1}{2}\|\mathbf{w}\|^2$. This maximizes $\frac{2}{\|w\|}$.

## Mathematical Programming Setting

Combining the above requirements we obtain

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

# Lagrange Function

THE AUSTRALIAN NATIONAL UNIVERSITY

## Objective Function

We have $\frac{1}{2}\|\mathbf{w}\|^2$.

## Constraints

Clearly the constraint

$$c_i(\mathbf{w}, b) := 1 - y_i(\langle w, \mathbf{x}_i \rangle + b) \leq 0$$

is a **convex** function. Hence we can use the default Lagrange approach.

## Lagrange Function

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \text{PrimalObjective} + \sum_i \alpha_i c_i \\ &= \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) \end{aligned}$$

## Saddle Point Condition

We need that the partial derivatives of $L$ with respect to $\mathbf{w}$ and $b$ vanish.

# Solving the Equations

THE AUSTRALIAN NATIONAL UNIVERSITY

## Lagrange Function

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b))$$

## Saddlepoint in w

$$\partial_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \iff \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

## Saddlepoint in $b$

$$\partial_b L(\mathbf{w}, b, \alpha) = -\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0$$

To obtain the dual optimization problem we have to substitute the values of $\mathbf{w}$ and $b$ into $L$. Note that the dual variables $\alpha_i$ have the constraint $\alpha_i \geq 0$.

# Solving the Equations



## Dual Optimization Problem

The terms linear in $\sum_i \alpha_i y_i$, i.e. the $b$-dependent term, vanishes.

$$\frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{m}\left[\alpha_i - \alpha_i y_i \sum_{j=1}^{m} \alpha_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right] = -\frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{m} \alpha_i$$

subject to $\sum_{i=1}^{m} \alpha_i y_i = 0$ and $\alpha_i \geq 0$ for all $1 \leq i \leq m$

## Practical Modification

We have to **maximize** the dual objective function. Typically we rewrite this as

$$\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{m} \alpha_i$$

$$\text{subject to} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } 1 \leq i \leq m$$

# Support Vector Expansion



## Solution in w

- $\mathbf{w}$ is given by a linear combination of training patterns $\mathbf{x}_i$ and the solution is **independent of the dimensionality of** $\mathcal{X}$.
- The expansion of $\mathbf{w}$ depends on the Lagrange multipliers $\alpha_i$.

## Kuhn-Tucker-Conditions

We know that at the optimal solution

$$\text{Constraint} \cdot \text{Lagrange Multiplier} = 0$$

In the present context this means that $\alpha_i(1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) = 0$. In other words $\alpha_i \neq 0$ implies that

$$y_i \left( \langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) = 1$$

**Only points at the decision boundary can contribute to the solution.**

This also allows us to compute $b$ via $b = y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle$.

# Kernels



## Nonlinearity via Feature Maps

In the linear optimization problem

$$\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{m} \alpha_i$$

$$\text{subject to} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } 1 \leq i \leq m$$

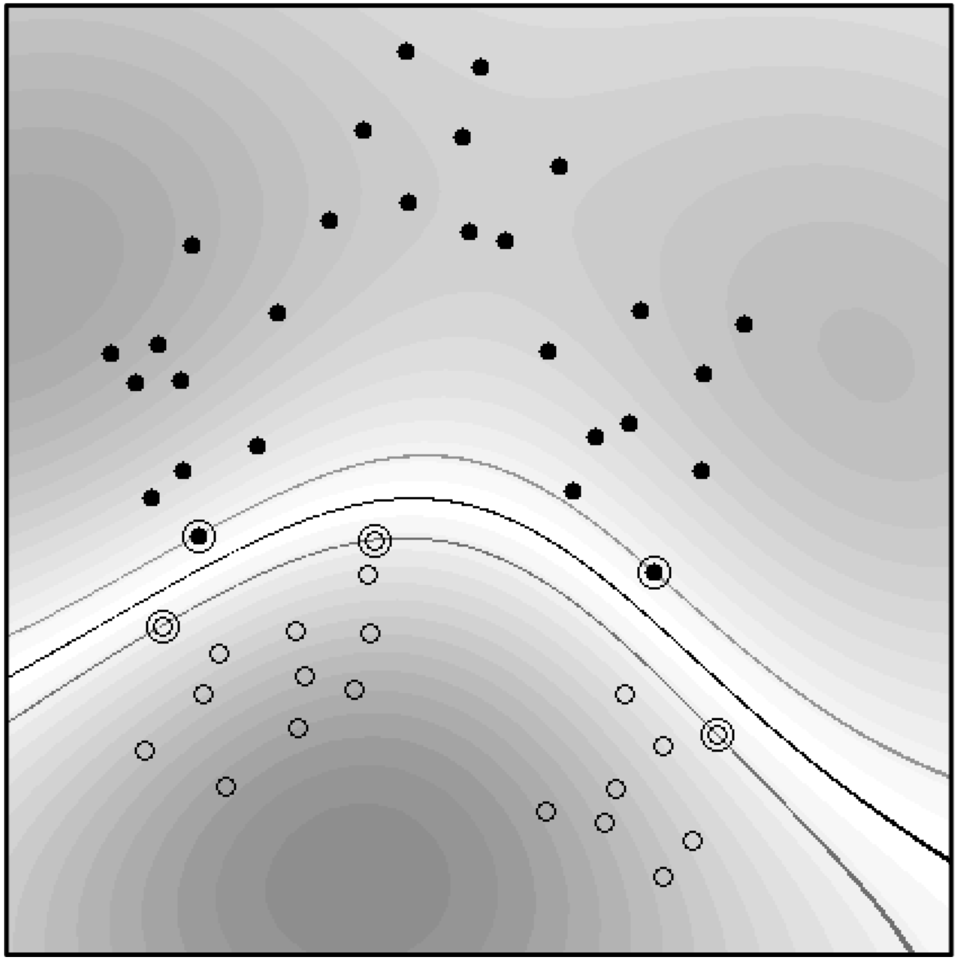
we replace $\mathbf{x}_i$ by $\Phi(\mathbf{x}_i)$ to obtain the new objective function

$$\text{minimize} \quad \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{m} \alpha_i$$

## Function Expansion

From $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i)$ we conclude $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b = \sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$.

# Examples and Problems



## Advantage

Works well when the data is noise free.

## Problem

Already a single wrong observation can ruin our estimate completely — we require that for all $i$ we have $y_i f(\mathbf{x}_i) \geq 1$.

## Idea

We have to limit the influence of individual observations by making the constraints less stringent (introduce slacks).

# Optimization Problem (Soft Margin)

THE AUSTRALIAN NATIONAL UNIVERSITY

## Recall: Hard Margin Problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(\langle w, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

## Softening the Constraints

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0 \text{ for all } 1 \leq i \leq m \end{aligned}$$

## Connection to Regularized Risk Functional

Up to scaling factors the margin term $\frac{1}{2}\|\mathbf{w}\|^2$ is the regularization term, the term in $\xi_i$ together with the constraints is the loss term, i.e.

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \frac{\lambda}{2}\|w\|^2$$

In our case $c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = \max(0, 1 - y_i f(\mathbf{x}_i))$.

# Lagrange Function and Constraints



## Lagrange Function

We have $m$ more constraints, namely those on the $\xi_i$, for which we will use $\eta_i$ as Lagrange multipliers.

$$L(\mathbf{w}, b, \xi, \alpha, \eta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left(1 - \xi_i - y_i(\langle\mathbf{w}, \mathbf{x}_i\rangle + b)\right) - \sum_{i=1}^{m}\eta_i\xi_i$$

## Saddle Point in $\mathbf{w}$

$$\partial_{\mathbf{w}} L(\mathbf{w}, b, \xi, \alpha, \eta) = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i = 0 \Longleftrightarrow \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i.$$

## Saddle Point in $b$

$$\partial_b L(\mathbf{w}, b, \xi, \alpha, \eta) = \sum_{i=1}^{m} -\alpha_i y_i = 0 \Longleftrightarrow \sum_{i=1}^{m}\alpha_i y_i = 0.$$

## Saddle Point in $\xi_i$

$$C - \alpha_i - \eta_i = 0 \Longleftrightarrow \alpha_i \in [0, C] \text{ with } \eta_i = C - \alpha_i.$$

# Dual Optimization Problem



The terms linear in $\sum_i \alpha_i y_i$, i.e. the $b$-dependent term, plus all the terms in $\xi_i$ and $\eta_i$ vanish. This is so since

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \eta) &= \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^m \xi_i(C - \alpha_i - \eta_i) + b\sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \alpha_i(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^m \alpha_i(1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle) \\ &= -\frac{1}{2}\sum_{i,j=1}^m \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle + \sum_{i=1}^m \alpha_i \end{aligned}$$

This is the dual objective function which will be **maximized** subject to

$$\sum_{i=1}^m \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \text{ for all } 1 \leq i \leq m.$$

The only difference to the unconstrained problem is that here $0 \leq \alpha_i \leq C$.

# SV Classification Machine

THE AUSTRALIAN NATIONAL UNIVERSITY