Support Vector Machine

Outline

Managerial motivation

Hard margin: the geometry

Soft margin

The dual formulation and the support vectors

Kernels: non-linear boundaries from a convex QP

Unbalanced classes and asymmetric costs

Support Vector Regression

Exercises

Learning objectives (1/2)

At the end of this chapter you will be able to:

- formulate hard and soft margin classification as QPs;
- derive and solve the dual, identifying the support vectors;

Learning objectives (2/2)

At the end of this chapter you will be able to:

- obtain non-linear boundaries with kernels while staying in the convex world;
- handle unbalanced classes with asymmetric costs and evaluate with the right metrics.

Prerequisites

- From this module, Chapter 2 «Background: convexity, quadratic programming and KKT conditions»: convex quadratic programs, the KKT conditions and the multipliers as shadow prices
- From the module *Convex Optimization*, Chapter 6 «Convex and concave optimization problems»: convex optimization problems and their optimality conditions
- From this module, Chapter 3 «The solver for nonlinear models»: quadratic objectives in `gurobipy`

Managerial motivation

Introduction: Managerial motivation

- The Support Vector Machine connects convex optimization directly to machine learning
- The classifier is obtained by solving a **QP**, with a clear geometric reading
- Immediate managerial applications: credit risk, churn, predictive maintenance, quality control, fraud

No machine learning libraries

- In this chapter we do *not* use ML libraries: every model is a QP written and solved with Gurobi
- The goal is to understand *what* a classifier optimizes
- **A modelling note**: the labels $y_i \in \{-1, +1\}$ are *observed data*, not variables
- The variables (coefficients, intercept, slacks, multipliers) are all continuous

The chapter closes the circle

- The same tools used for production, logistics and finance — convex QPs, duality, multipliers — are the engine of one of the most elegant classifiers of machine learning
- Understanding the SVM as an optimization problem, and not as a black box from a library, gives full control over what the model is really doing

Summary

- The SVM is a classifier obtained by solving a convex QP
- The labels y_i are data; all the variables are continuous
- Every model of the chapter is written and solved with Gurobi, without ML libraries

Hard margin: the geometry

Introduction: Hard margin

- We start from the simplest case: two classes separable by a hyperplane
- Among all separating hyperplanes we look for the one with the widest **margin**
- We will see that maximizing the margin is a convex QP with linear constraints

The problem in words

- *We decide* the coefficients of a hyperplane that separates two classes of observations
- *The objective* is to maximize the margin, that is, the distance between the hyperplane and the closest points of the two classes (a classifier with a wide margin generalizes better)
- *The constraints* require every observation to lie on the right side of the hyperplane — in the soft margin, up to a slack paid in the objective

Data (input of the model) (1/2)

Symbol	Type	Meaning
n	$\in \mathbb{Z}_{\geq 1}$	number of observations of the training set, indexed by $i \in \{1, 2, \ldots, n\}$
p	$\in \mathbb{Z}_{\geq 1}$	number of features, indexed by $j \in \{1, 2, \ldots, p\}$
x_{ij}	$\in \mathbb{Q}$	value of feature j for observation i; $\boldsymbol{x}_i \in \mathbb{Q}^p$ is the vector of observation i

Data (input of the model) (2/2)

Symbol	Type	Meaning
y_i	$\in \{-1, +1\}$	*observed* class label of i (a datum, not a variable)
C	$\in \mathbb{Q}_{>0}$	weight of the margin violations (soft margin only)

- C is upper case to follow the machine learning literature: an exception to the convention that reserves capital letters for sets and matrices

Decision variables

We introduce the following $p + 1$ free variables:

$$\begin{cases} w_j = \text{coefficient } j \text{ of the separating hyperplane} \\ b = \text{intercept of the hyperplane} \end{cases}$$

$$\forall j \in \{1, 2, \ldots, p\}$$

They define the hyperplane

$$\sum_{j=1}^{p} w_j \, x_j + b = 0$$

Hard margin QP model

Hard margin QP model

$$\min \quad \frac{1}{2}\sum_{j=1}^{p} w_j^2 \tag{12.1a}$$

$$\text{subject to} \quad y_i\left(\sum_{j=1}^{p} w_j\, x_{ij} + b\right) \geq 1, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{12.1b}$$

$$w_j \gtreqless 0, \quad \forall j \in \{1, 2, \ldots, p\}, \tag{12.1c}$$

$$b \gtreqless 0. \tag{12.1d}$$

Description of the hard margin QP model

- The convex quadratic objective function minimizes $\frac{1}{2}\|\boldsymbol{w}\|^2$: since the geometric margin between the supporting hyperplanes is $2/\|\boldsymbol{w}\|$, minimizing it is equivalent to *maximizing the margin*
- The linear classification constraints impose that every observation lies on the right side of the hyperplane, at a distance of at least one unit in the scale of $\boldsymbol{w}$ (n linear constraints)
- The last constraints define the variables of the model, all free

Example: Two points, everything by hand (1/3)

- $\boldsymbol{x}_1 = (0, 0)$ with $y_1 = -1$ and $\boldsymbol{x}_2 = (2, 2)$ with $y_2 = +1$

Step 1 — constraints

$$-(b) \geq 1 \qquad \text{and} \qquad 2w_1 + 2w_2 + b \geq 1$$

that is,

$$b \leq -1 \qquad \text{and} \qquad 2(w_1 + w_2) \geq 1 - b \geq 2$$

Example: Two points, everything by hand (2/3)

Step 2 — symmetry and optimum

- To minimize $w_1^2 + w_2^2$ with $w_1 + w_2 \geq 1$ it is best to take $w_1 = w_2 = t$ and the inequalities active:

$$4t + b = 1, \quad b = -1 \quad \Rightarrow \quad t = 1/2$$

- Hence $\boldsymbol{w} = (0.5; 0.5)$ and $b = -1$
- The hyperplane $0.5x_1 + 0.5x_2 = 1$ is the perpendicular bisector of the segment between the two points

Example: Two points, everything by hand (3/3)

Step 3 — margin

$$\frac{2}{\|\boldsymbol{w}\|} = \frac{2}{\sqrt{0.5}} = 2\sqrt{2} = 2.83$$

- It is exactly the distance between the two points
- Both constraints are active: both points are *support vectors*
- With real overlapping data, however, the hard margin is infeasible: the soft margin is needed

Summary

- The hard margin looks for the separating hyperplane of maximum margin
- Margin $= 2/\|\boldsymbol{w}\|$: maximizing it is equivalent to minimizing $\frac{1}{2}\|\boldsymbol{w}\|^2$
- It is a convex QP with n linear constraints and $p + 1$ free variables
- The points with an active constraint are the *support vectors*
- With overlapping classes the model is infeasible

Soft margin

Introduction: Soft margin

- With real overlapping data the hard margin is infeasible
- We allow some observations to violate the margin, paying the violation in the objective
- The parameter C governs the trade-off between a wide margin and fit to the data

Slack variables

One adds n non-negative slack variables:

$$\xi_i = \text{margin violation of observation } i, \quad \forall i \in \{1, 2, \ldots, n\}$$

- $\xi_i = 0$: the observation respects the margin
- $\xi_i > 1$: the observation is misclassified

Soft margin QP model (primal)

Soft margin QP model (primal)

$$\min \quad \frac{1}{2}\sum_{j=1}^{p} w_j^2 + C\sum_{i=1}^{n}\xi_i \tag{12.2a}$$

$$\text{subject to} \quad y_i\Big(\sum_{j=1}^{p} w_j\, x_{ij} + b\Big) \geq 1 - \xi_i, \quad \forall i \in \{1,2,\ldots,n\}, \tag{12.2b}$$

$$w_j \gtreqless 0, \qquad \forall j \in \{1,2,\ldots,p\}, \tag{12.2c}$$

$$b \gtreqless 0, \tag{12.2d}$$

$$\xi_i \geq 0, \qquad \forall i \in \{1,2,\ldots,n\}. \tag{12.2e}$$

Description of the soft margin QP model

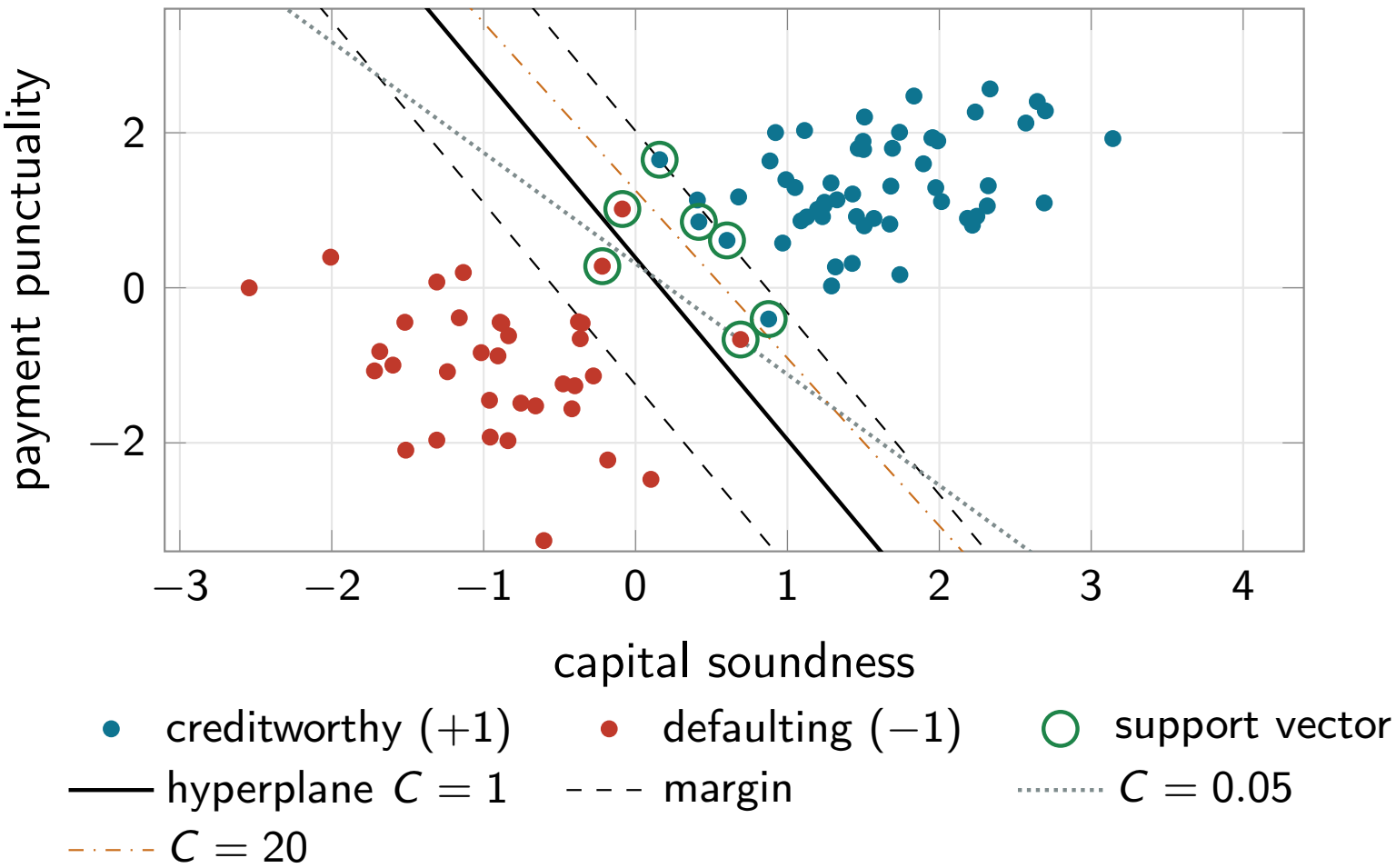
- The convex quadratic objective function balances two terms:
  - the simplicity of the model ($\frac{1}{2}\|\boldsymbol{w}\|^2$, wide margin)
  - the cost of the violations on the training set ($C \sum_i \xi_i$)
- C is the price assigned to the violations
- The linear constraints require correct classification *up to the slack* ξ_i, paid in the objective (n linear constraints)
- Variables: w_j and b free, ξ_i non-negative

Case study: credit risk

- 90 customers, two standardized indicators: capital soundness and punctuality of payments
- 55 creditworthy and 35 defaulting, partially overlapping

C	$\boldsymbol{w}$	b	margin	errors	in the margin
0.05	(0.678; 0.474)	−0.148	2.418	2	22
1.00	(1.433; 0.611)	−0.236	1.284	2	7
20.00	(5.058; 2.338)	−2.936	0.359	0	3

Soft margin: hyperplane, margin, support vectors



How to read the chart

- The 90 customers in the plane of the two indicators: creditworthy (teal), defaulting (red), support vectors circled in green
- The black line is the optimal hyperplane for $C = 1$ with its margin lines (dashed)
- The grey and orange lines are the hyperplanes for $C = 0.05$ (wide margin) and $C = 20$ (narrow margin)

The regularization–fit trade-off

- A small C buys a wide margin while accepting 22 points inside the margin
- A large C chases every point of the training set (zero errors, a very narrow margin)
- This is the **regularization–fit** trade-off: C must be chosen with validation data, never by looking at the test set
- The three lines are similar but not equal: the slope changes because the "paid for" points change

Summary

- The soft margin adds the slacks $\xi_i \geq 0$, each paid C in the objective
- It remains a convex QP, always feasible
- Small C: wide margin, more violations; large C: narrow margin, fit to the training set
- C is chosen with validation data, never with the test set

The dual formulation and the support vectors

Introduction: The dual formulation

- We associate a multiplier with every classification constraint and write the dual of the soft margin
- The dual reveals which observations really determine the boundary: the **support vectors**
- The dual depends on the data only through inner products: it is the door to kernels

Multipliers

By associating a multiplier α_i with each classification constraint of the soft margin, one obtains n new variables:

$$\alpha_i = \text{multiplier of the classification constraint of observation } i$$

$$\forall i \in \{1, 2, \ldots, n\}$$

Dual QP model of the soft margin

Dual QP model of the soft margin

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, \boldsymbol{x}_i' \boldsymbol{x}_j \tag{12.3a}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0, \tag{12.3b}$$

$$\alpha_i \leq C, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{12.3c}$$

$$\alpha_i \geq 0, \quad \forall i \in \{1, 2, \ldots, n\}. \tag{12.3d}$$

Description of the dual model

- The concave quadratic objective function depends on the data *only* through the inner products $\boldsymbol{x}_i'\boldsymbol{x}_j$: this is the observation that opens the door to the kernel
- The linear constraint $\sum_i \alpha_i y_i = 0$ follows from stationarity with respect to the intercept b (one linear constraint)
- The constraints $\alpha_i \leq C$ bound each multiplier by the price C of the violations (n linear constraints)
- The constraints $\alpha_i \geq 0$ define the variables of the model

Primal-dual relation

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \, y_i \, \boldsymbol{x}_i \qquad\qquad f(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}'\boldsymbol{x} + b)$$

On the credit risk case ($C = 1$):

- $\boldsymbol{w}$ from the dual $= (1.433;\ 0.611)$, $b = -0.236$: identical to the primal
- dual value 4.7868 = primal value (strong duality)

multiplier	points	meaning
$\alpha_i = 0$	83	do NOT affect the boundary
$0 < \alpha_i < C$	2	exactly on the margin
$\alpha_i = C$	5	inside the margin or misclassified

Why they are called support vectors (1/2)

- The solution depends *only* on the 7 points with $\alpha_i > 0$: it is they that "support" the hyperplane
- The other 83 could disappear without changing anything
- Moreover, the dual depends on the data only through the inner products $\boldsymbol{x}_i'\boldsymbol{x}_j$: this observation opens the door to the kernel

Why they are called support vectors (2/2)

The KKT complementarity organizes everything:

- $\alpha_i = 0 \Rightarrow$ point beyond the margin
- $0 < \alpha_i < C \Rightarrow$ exactly on the margin (from these one recovers b)
- $\alpha_i = C \Rightarrow$ the constraint "pays" a slack $\xi_i > 0$

Summary

- The dual has one variable $\alpha_i \in [0, C]$ per observation
- $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$: the boundary depends only on the points with $\alpha_i > 0$, the support vectors
- Strong duality: dual value = primal value
- The data enter the dual only through the inner products $\boldsymbol{x}_i' \boldsymbol{x}_j$

Kernels: non-linear boundaries from a convex QP

Introduction: Kernels

- Not all classes can be separated by a straight line
- Exploiting the dual, we replace the inner product with a **kernel**
- We obtain curved boundaries while staying in a convex QP

The RBF kernel

- We replace $\boldsymbol{x}_i'\boldsymbol{x}_j$ with a kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for example the Gaussian RBF (*Radial Basis Function*) kernel

$$K(\boldsymbol{x}, \boldsymbol{z}) = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{z}\|^2}$$

- One classifies in a transformed space without ever building it:

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right)$$

- The boundary can be arbitrarily curved, but training *remains a convex QP* if the kernel is valid (positive semidefinite matrix)

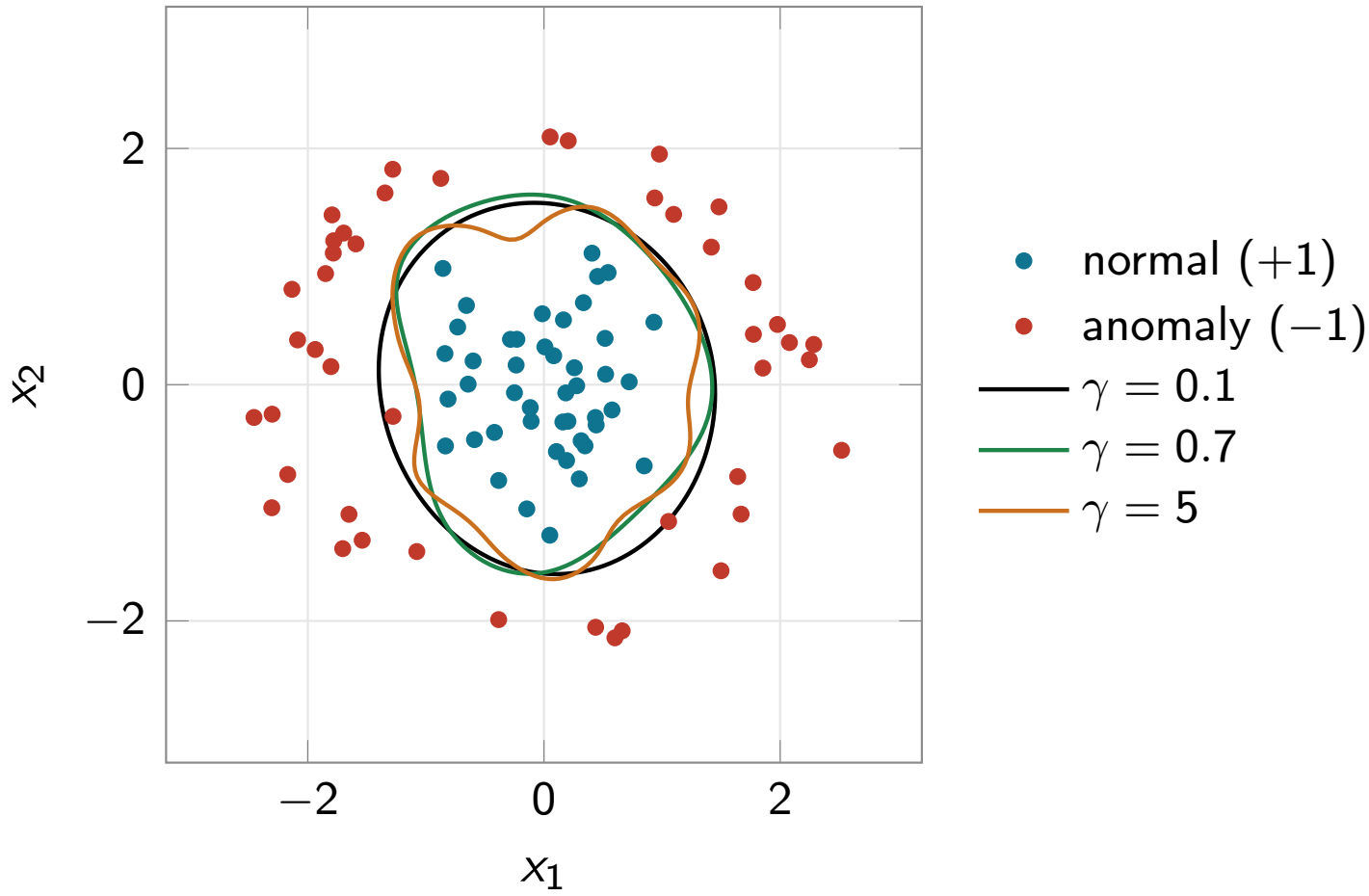
Case study: "ring-shaped" anomalies

- 90 process observations
- Normal operation in the centre, the anomalies arranged in a ring
- No straight line can separate them

γ	training errors	support vectors
0.1	1	20
0.7	0	15
5.0	0	61

With $\gamma = 5$ the model "memorizes" the points.

RBF kernel: the level-0 boundary



How to read the chart

- The "ring-shaped" dataset: normal operation in the centre, anomalies in a ring
- The curves are the decision boundaries of the RBF kernel for three values of γ
- With $\gamma = 0.1$ the boundary is a regular circle
- With $\gamma = 5$ it breaks up around the individual points (overfitting)

Overfitting

Observation 1: Convexity does not mean generalization

With $\gamma = 5$ the boundary breaks up into islands around the individual points and the support vectors rise to 61 out of 90: the model is memorizing the training set (overfitting), while still remaining a convex QP.

Convexity of training and the ability to generalize are different properties. C and γ must be chosen together by cross-validation.

Summary

- The kernel replaces the inner product in the dual: non-linear boundaries without building the transformed space
- With a valid kernel training remains a convex QP
- Large γ: fragmented boundary, many support vectors, overfitting
- C and γ are chosen together by cross-validation

Unbalanced classes and asymmetric costs

Introduction: Unbalanced classes

- In managerial applications the important class is rare
- An error on the rare cases costs more (an undetected defaulter, an undiscovered fraud)
- We see how to build asymmetric costs into the model

Weighting the slacks

It is enough to weight the slacks differently for the two classes:

$$C_- \sum_{i:\, y_i=-1} \xi_i \; + \; C_+ \sum_{i:\, y_i=+1} \xi_i$$

	precision	recall defaulters
equal costs ($C = 1$)	1.00	0.94
$10\times$ cost on the defaulters	0.97	1.00

Decision costs, not frequencies

- With the 10× cost the boundary moves back towards the "healthy" class
- No defaulter escapes (recall 100%) at the price of a few more false alarms (precision from 100% to 97%)
- The weights must reflect the *decision costs*: how much a loan granted to someone who will not repay costs vs. a good customer turned down
- Not the frequencies of the classes

Summary

- With unbalanced classes one uses two prices C_- and C_+ for the slacks
- The model remains a convex QP
- One evaluates with precision and recall, not with accuracy alone
- The weights reflect the decision costs, not the frequencies

Support Vector Regression

Introduction: Support Vector Regression

- The same idea applies to estimating a continuous function
- SVR ignores the errors within a tube of width ε
- Only the points outside the tube determine the estimate

The SVR model

$$\min \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

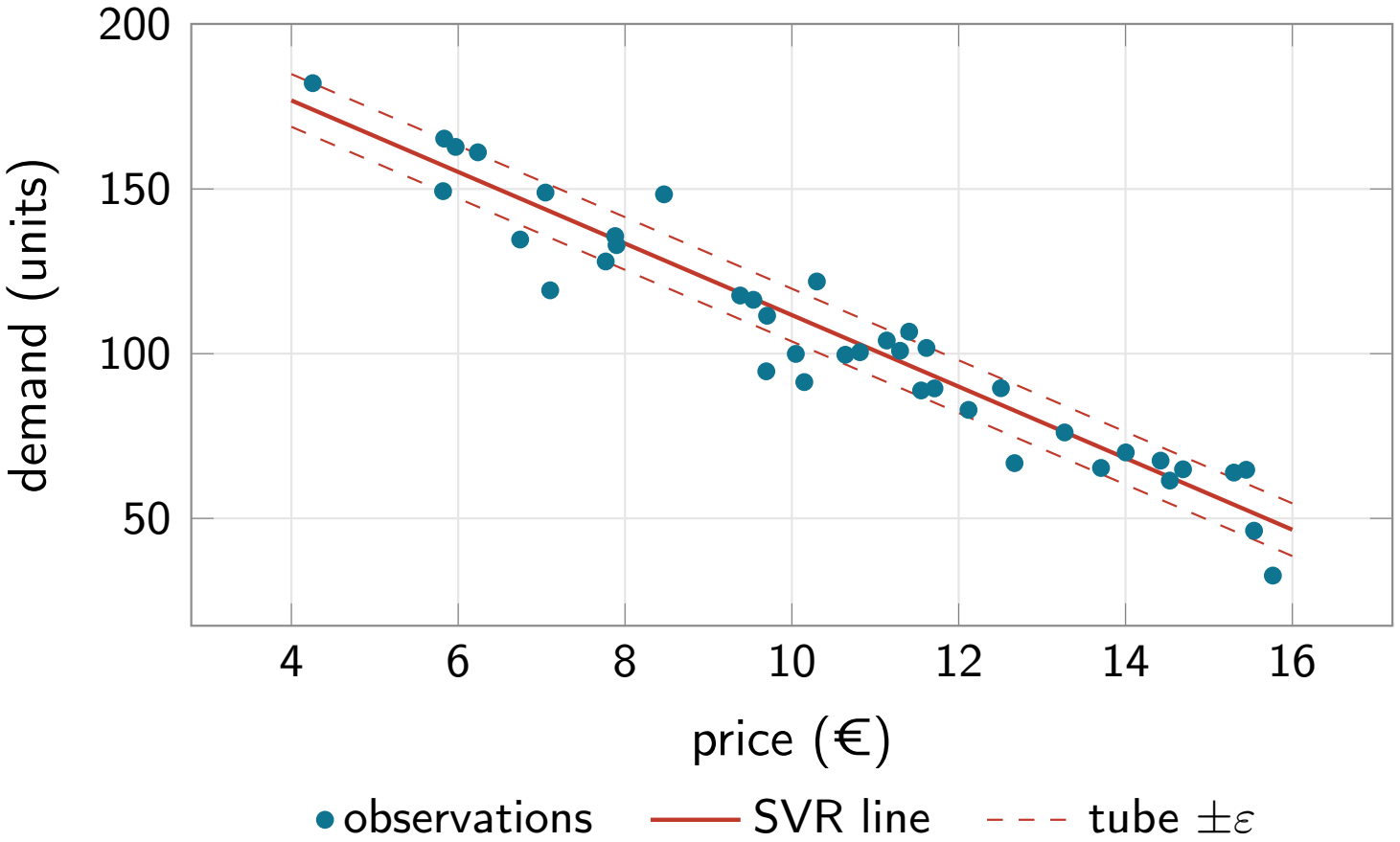
with

$$|y_i - (\boldsymbol{w}'\boldsymbol{x}_i + b)| \le \varepsilon + \xi_i^{(*)} \qquad \text{for every } i$$

On the price–demand relation (40 observations):

- SVR line: demand $= -10.86 \cdot \text{price} + 220.32$ (true: $-11p + 220$)
- tube $\varepsilon = 8$: 14 points out of 40 outside the tube $\rightarrow$ only they determine the line

SVR: only the points outside the tube pay a penalty



How to read the chart

- The 40 observations of the price–demand relation
- The estimated line and the tolerance tube $\pm\varepsilon = 8$
- Only the 14 points outside the tube pay a penalty and determine the line

Summary

- SVR estimates a continuous function with a convex QP
- Errors within the tube $\pm\varepsilon$ are not paid
- Only the points outside the tube determine the line: they are the support vectors of regression

Exercises

Introduction: Exercises

- The exercises put the concepts of the chapter to the test
- They are solved by hand or with the lab script and Gurobi
- They cover hard margin, complementarity, the choice of C, the decision threshold and the One-Class SVM

Exercise: check

- Solve by hand the hard margin with

$$\boldsymbol{x}_1 = (0, 0),\ y_1 = -1 \qquad \text{and} \qquad \boldsymbol{x}_2 = (4, 0),\ y_2 = +1$$

- Check with Gurobi:

$$\boldsymbol{w} = (0.5;\ 0), \qquad b = -1, \qquad \text{margin } 4$$

Exercise: complementarity

- For $C = 1$, list from the script the 7 points with $\alpha_i > 0$
- Check numerically the three complementarity conditions:

$$\alpha_i = 0 \Rightarrow \xi_i = 0 \text{ and margin } \geq 1; \quad \text{and so on}$$

Exercise: validation curve

- Split the 90 customers into training (60) and validation (30)
- Plot the validation error for $C \in \{0.01; 0.1; 1; 10; 100\}$ and choose C
- Is the training error monotone in C? And the validation error?

Exercise: decision threshold

- For the same model ($C = 1$), move the threshold: classify as defaulting if

$$\boldsymbol{w}'\boldsymbol{x} + b < \theta, \qquad \theta \in [-1, 1]$$

- Plot precision and recall as a function of θ
- Where is the point that maximizes profit if an undetected defaulter costs 10 and a false alarm 1?

Exercise: one-class

- Implement the One-Class SVM on the ring-shaped dataset using only the normal points:

$$\min \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \quad \text{subject to} \quad \boldsymbol{w}'\phi(\boldsymbol{x}_i) \geq \rho - \xi_i$$

- RBF kernel, $\gamma = 0.7$
- Which fraction of anomalies does it detect for $\nu \in \{0.05; 0.1; 0.2\}$?

Final summary (1/2)

Models:

- Hard margin: convex QP, maximizes the margin $2/\|\boldsymbol{w}\|$
- Soft margin: slacks ξ_i paid C, always feasible
- Dual: $\alpha_i \in [0, C]$, $\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i$

Support vectors:

- Only the points with $\alpha_i > 0$ determine the boundary
- KKT complementarity says where each point lies

Final summary (2/2)

Extensions:

- Kernels: non-linear boundaries from a convex QP
- Asymmetric costs C_-, C_+ for unbalanced classes
- SVR: regression with a tolerance tube $\pm\varepsilon$

Modeller's choices:

- C and γ with validation data, never with the test set
- The weights reflect the decision costs, not the frequencies