

Advanced Topics in Machine Learning (2010, Part II)

1. Review of Regularisation

Massimiliano Pontil

Assessment

1. Written Examination (2.5 hours, 50%)
2. Coursework Section (2 pieces, 50%)
(deliver on-time, penalty otherwise)
 - To pass this course, students must:
Obtain an average of at least 50% when the coursework and exam components of the course are weighted together
 - The examination rubric is: Answer TWO questions out of FOUR. All questions carry equal marks.

Material

- Lecture notes
 - <http://www.cs.ucl.ac.uk/staff/m.pontil/courses/index-ATML10.htm>
- Reference Books
 - N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines Cambridge University Press, 2001.
 - J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
 - B.Scholkopf and A.J. Smola. Learning with Kernels. MIT Press, 2002.
 - S. Boyd and L. Vandenberghe. Convex Optimization, Cambridge University Press, <http://www.stanford.edu/boyd/cvxbook/>

Prerequisites

- Calculus (real-valued functions, limits, derivatives ...)
- Fundamentals of linear algebra (vectors, matrices, eigenvectors/eigenvalues ...),
- A bit of optimization theory (Lagrange multipliers)
- Supervised Learning (GI01 / COMPM055)

Course Outline

- **Lecture 1** Regularization review: ridge regression, SVMs, SVM dual
- **Lecture 2** Representer Theorem; obtaining a problem of limited dimensionality; kernels and examples
- **Lecture 3** Convex optimization: convex sets, convex functions, determining convexity; types of convex programs and algorithms; convex learning problems
- **Lecture 4** Learning the kernel: combining kernels, selection of hyperparameters; general formulation and proof of convexity; SVM, ridge regression cases; algorithms

Course Outline (contd.)

- **Lecture 5** Multi-task learning: definition and applications; trace norm regularization and algorithms; relation to matrix factorization; multi-task representer theorem; other methods in machine learning

Today's Plan

- Ridge regression
- SVM classification
- Regularization in general
- Ridge regression and SVM as special cases of regularization

Notation

We summarize our notation. More symbols will be introduced during the course when needed.

$\mathbb{N}$ is the set of natural numbers, $\mathbb{R}$ the set of real ones

$\mathbb{R}^d$ denotes the set of d –dimensional vectors. $x \in \mathbb{R}^d$ is regarded as a column vector ($d \times 1$ matrix), $x^\top$ is its transpose ($1 \times d$ matrix)

$\mathbf{M}_{d,n}$ denotes the set of $d \times n$ matrices. If $A = (A_{ij})_{\substack{i=1,\dots,d \\ j=1,\dots,n}}$ is a $d \times n$ matrix, $A^\top$ denotes its transpose and A^+ its pseudoinverse.

I_d denotes the $d \times d$ identity matrix

Supervised Learning Problem

- Given a set of **input / output** pairs (**training set**) we wish to compute the functional relationship between the input and the output

$$x \xrightarrow{f} y$$

- For simplicity we will be assuming, throughout the course, that $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$
- Ex. 1 (people detection): given an image we wish to say if it depicts a person or not. The output is one of 2 possible categories
- Ex. 2 (pose estimation): we wish to predict the pose of a face image. The output is a real number describing the face rotation angle.

In both problems the input is a high dimensional vector x representing pixel intensity/color

Linear Regression

- We wish to learn a function $f(x) = w^\top x$ which best interpolates a data set $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \subseteq \mathbb{R}^d \times \mathbb{R}$
 - If the data have been generated as $y_i = w^{*\top} x_i$ (i.e. there is no noise), the inputs x_i are linearly independent and $m = d$ then there is a unique interpolant w such that

$$X^\top w = y$$

where $X = \begin{pmatrix} | & & | \\ x_1 & \dots & x_m \\ | & & | \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$.

- If $m < d$, this problem is *ill-posed*

Linear Regression (contd.)

Remark: If we wish to learn a non-homogeneous function $f(x) = w^\top x + b$ (i.e. there is bias) we simply add one more dimension to x and w . Specifically, we replace x by the vector $\begin{pmatrix} x \\ 1 \end{pmatrix}$ and w by $\begin{pmatrix} w \\ b \end{pmatrix}$.

Ridge Regression

- *Regularization* is an effective means to make the learning problem well-posed
- We solve the optimization problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m (y_i - w^\top x_i)^2 + \gamma \|w\|^2 \quad (1)$$

The scalar parameter $\gamma > 0$ defines a trade-off between the error on the data and the norm of the vector w (degree of regularization)

Ridge Regression (contd.)

We use $\|\cdot\|$ to denote the L_2 norm $\|\cdot\|_2$, i.e.

$$\|w\|_2 = \left(\sum_{\ell=1}^d w_{\ell}^2 \right)^{\frac{1}{2}}$$

where w_{ℓ} denotes the ℓ -th component of vector w .

Regularization with the L_2 norm (the vector length) of w favors “flatter” linear functions f .

Ridge Regression (contd.)

Write the objective in (1) as

$$\begin{aligned} \|y - X^\top w\|^2 + \gamma \|w\|^2 &= \\ (y - X^\top w)^\top (y - X^\top w) + \gamma w^\top w \end{aligned}$$

Setting the gradient to zero

$$2XX^\top w - 2Xy + 2\gamma w = 0$$

we obtain the *regularized solution*

$$w = (XX^\top + \gamma I_d)^{-1} Xy$$

Singular Value Decomposition (review)

Let $Z \in \mathbf{M}_{d,n}$ be a $d \times n$ matrix. Singular value decomposition (SVD) establishes that

$$Z = \underset{d \times r}{U} \underset{r \times r}{\Sigma} \underset{r \times n}{V^\top}$$

where U and V are (rectangular) orthogonal matrices ($U^\top U = V^\top V = I_r$) and $\Sigma = \text{Diag}(\sigma_1, \dots, \sigma_r)$, $\sigma_1, \dots, \sigma_r > 0$

$r \leq \min\{d, n\}$ is the *rank* of Z

Alternatively, we can write

$$Z = \sum_{i=1}^r \sigma_i u_i v_i^\top$$

where $U = (u_1, \dots, u_r)$, $V = (v_1, \dots, v_r)$

Singular Value Decomposition (review)

$\sigma_1, \dots, \sigma_r$ are called the *singular values* of Z

Moreover,

$$ZZ^\top = U\Sigma^2U^\top, \quad Z^\top Z = V\Sigma^2V^\top$$

and $\text{rank}(Z) = \text{rank}(ZZ^\top) = \text{rank}(Z^\top Z)$

The *pseudoinverse* of Z is the matrix

$$Z^+ := V\Sigma^{-1}U^\top = \sum_{i=1}^r \sigma_i^{-1} v_i u_i^\top$$

If $Z \in \mathbf{M}_{d,d}$ is invertible, $Z^+ = Z^{-1}$.

(See also Appendix A.5.4 in [Boyd] <http://www.stanford.edu/boyd/cvxbook/>)

Minimal Norm Interpolation

- As γ goes to zero the solution $\hat{w}$ of (1) tends to

$$w_0 := (XX^\top)^+ Xy = (X^\top)^+ y \quad (4)$$

- In the case when the linear system

$$X^\top w = y$$

has a solution, (4) is the solution with the smallest norm $\|w\|$

This is an example of *minimal norm interpolation*

Support Vector Machines (SVMs)

- Consider the problem of *classification*
- Find optimal separating hyperplane; data are not necessarily linearly separable (*aka* soft-margin linear SVM)
- For now, use linear functions only (no kernels)
- Connect to regularization

Optimal Hyperplane

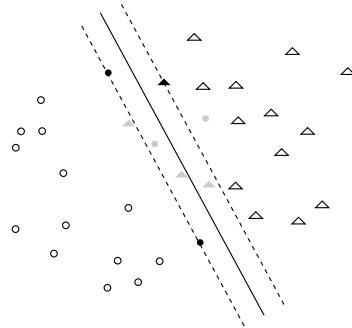
Training set $S = \{(x_1, y_1), \dots, (x_m, y_m)\} \subseteq \mathbb{R}^d \times \{-1, 1\}$

We consider hyperplanes containing the origin, i.e. of the form
 $H_w = \{x \in \mathbb{R}^d : w^\top x = 0\}$

Remark: As we did before, if we wish to learn a non-homogeneous function $f(x) = w^\top x + b$ we simply add one more dimension to x and w . So, in the following presentation the bias b is ignored. Because of this, the hyperplanes considered go through the origin.

We do **not** assume that the data are linearly separable

Optimal Hyperplane (contd.)



The distance $\rho_x(w)$ of a point x from a hyperplane H_w is

$$\rho_x(w) := \frac{|w^\top x|}{\|w\|}$$

Introduce slack variables $\xi_i \geq 0$, $i = 1, \dots, m$, to relax the separation constraints

Optimal Hyperplane (contd.)

Rescaling w with a positive constant gives the same hyperplane, hence the *margin* of H (ignoring the shaded data points) remains the same

Canonical hyperplane: scale $\|w\|$ so that $\min_{i:\xi_i=0} y_i(w^\top x_i) = 1$
(a data-dependent parameterization)

That is, $\xi_i > 0 \iff y_i(w^\top x_i) < 1$

We wish to maximize the margin $\frac{1}{\|w\|}$ subject to a bound on $\sum_{i=1}^m \xi_i$

Optimal Hyperplane (contd.)

$$\begin{aligned} & \max_w \left\{ \frac{1}{\|w\|} : y_i (w^\top x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \sum_{i=1}^m \xi_i \leq U \right\} \\ &= \frac{1}{\min_w \left\{ \|w\| : y_i (w^\top x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \sum_{i=1}^m \xi_i \leq U \right\}} \end{aligned}$$

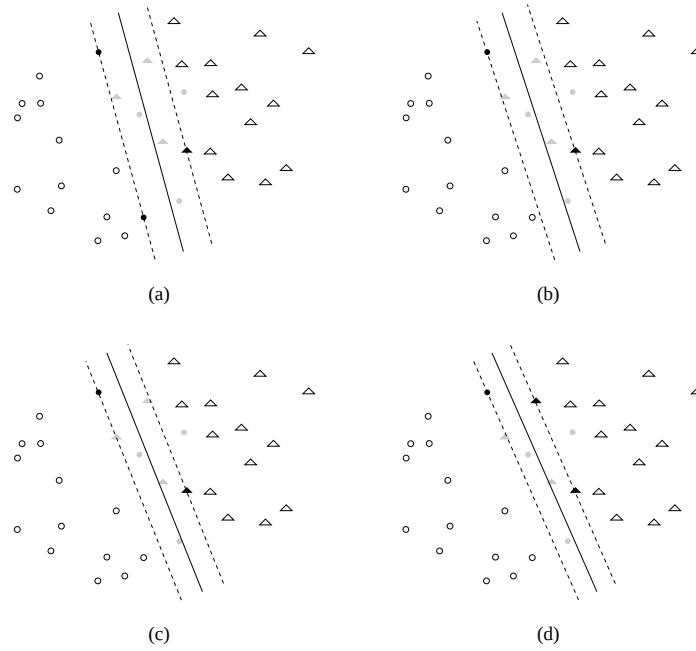
SVM Problem (Primal)

Problem **(P)**

Minimize $\frac{1}{2}\|w\|^2 + C \sum_{i=1}^m \xi_i$

subject to $y_i(w^\top x_i) \geq 1 - \xi_i,$
 $\xi_i \geq 0, \quad i = 1, \dots, m$

The role of the parameter C



Optimal separating hyperplane for four increasing values of C . Increasing C decreases the margin while decreasing the training error $\sum_{i=1}^m \xi_i$.

Saddle Point

The solution of problem **(P)** is equivalent to determining the **saddle point** of the Lagrangian function

$$L(w, \xi, \alpha, \beta) = \frac{1}{2}w^\top w + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i(w^\top x_i) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i \quad (5)$$

where $\alpha_i \geq 0, \beta_i \geq 0$ are the Lagrange multipliers.

L has a minimum wrt. w, ξ and a maximum wrt. α, β .

Karush-Kuhn-Tucker Conditions

Differentiating wrt. w we obtain:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \quad \implies \quad w = \sum_{i=1}^m \alpha_i y_i x_i \quad (6)$$

Differentiating wrt. ξ_i we obtain:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \quad (7)$$

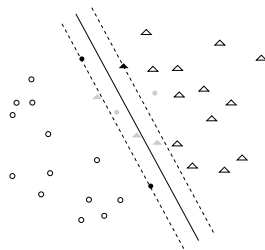
Karush-Kuhn-Tucker Conditions (contd.)

$$\alpha_i [y_i(w^\top x_i) - 1 + \xi_i] = 0 \quad (*)$$

$$(C - \alpha_i)\xi_i \stackrel{(7)}{=} \beta_i \xi_i = 0 \quad (**)$$

(complementary slackness)

The sample points x_i for which $\hat{\alpha}_i > 0$ are called **support vectors**



A Closer Look at the KKT Conditions

Equations (*) (**) imply that

- $y_i(\hat{w}^\top x_i) > 1 \implies \hat{\alpha}_i = 0$ (not a SV)
(used the second constraint in **(P)**)
- $y_i(\hat{w}^\top x_i) = 1 \implies \hat{\alpha}_i \in [0, C]$ (if $\hat{\alpha}_i > 0$, a SV 'on the margin')
- $y_i(\hat{w}^\top x_i) < 1 \implies \hat{\alpha}_i = C$ (a SV with positive slack ξ_i)
(used the first constraint in **(P)**)

Remark: Conversely, from eqs. (*),(**), if $\hat{\alpha}_i = 0$ then $y_i(\hat{w}^\top x_i) \geq 1$, $\xi_i = 0$; if $\hat{\alpha}_i \in (0, C)$ then $y_i(\hat{w}^\top x_i) = 1$, $\xi_i = 0$; if $\hat{\alpha}_i = C$ then $y_i(\hat{w}^\top x_i) \leq 1$, $\xi_i \geq 0$.

Dual Problem

Substituting eqs.(6) and (7) in eq.(5) leads to the **dual problem**

Problem **(D)**

$$\text{Maximize} \quad -\frac{1}{2} \alpha^\top A \alpha + \sum_{i=1}^m \alpha_i$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m$$

where A is the $m \times m$ matrix $A = \left(y_i y_j x_i^\top x_j \right)_{i,j=1}^m$

Note the “box constraints” on α

KKT Conditions and Support Vectors

Important note: the complexity of problem **(D)** depends on the sample size m , not on the number d of input components

If $\hat{\alpha}$ is a solution of the dual problem **(D)** then, by (6), the solution $\hat{w}$ of the primal problem **(P)** is given by

$$\hat{w} = \sum_{i=1}^m \hat{\alpha}_i y_i x_i$$

(as we shall see later, this is the *Representer Theorem*)

Note that $\hat{w}$ is a linear combination of only the x_i for which $\hat{\alpha}_i > 0$ (the **support vectors**)

Some Remarks on SVMs

- The fact that the optimal separating hyperplane is determined only by the support vectors is remarkable. Often, the support vectors are a small subset of the training data
- All the information contained in the data set is summarized by the support vectors: the whole data set could be replaced by only these points and the **same** hyperplane would be found
- A new point x is classified as $\text{sgn} \left(\sum_{i=1}^m \hat{\alpha}_i x_i^\top x \right)$

SVM Computational Issues

- Problems **(P)** and **(D)** are *quadratic programs (QP)* (as we shall see later in the course). Several methods from convex optimization exist for solving QP problems
- Very often, the dimensionality d is much larger than the number of examples m . In such cases, it is preferable to solve the dual problem
- In some data mining applications, instead, $m \gg d$ and it is more efficient to solve the primal problem

Regularization

- *Regularization* is a general approach for machine learning, but also for many other problems in science and engineering
- Long history: motivated by ill-posed problems in the 60s (Tykhonov and others)
- We shall see that ridge regression and SVMs are special cases of the regularization framework

Regularization (contd.)

Solve the optimization problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m E(w^\top x_i, y_i) + \gamma \Omega(w)$$

- $E : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is called the *loss function*
- $\Omega : \mathbb{R}^d \rightarrow \mathbb{R}$ is called the *regularizer*
- $\gamma > 0$ is called the *regularization parameter*

Regularization (contd.)

We shall focus on the problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m E(w^\top x_i, y_i) + \gamma \|w\|^2 \quad (8)$$

- The loss function E measures the error on the training data
A common choice is *square loss*, $E(a, b) = (a - b)^2$
- The regularizer (square of the L_2 norm) favors simpler solutions (having small norm)
- Usually, E is chosen to be *convex* (defined later in the course)

Effect of the Regularization Parameter

- Regularization *trades off* between data fit (E term) and regularity (Ω term)

If γ is large, the solution tends to be simpler

If γ is small, the solution tends to fit the data better

Cross Validation

- How do we select γ in practice?

- Randomly split the training data into $\mathcal{T} = \{(x_i, y_i) : i \in I\}$ and $\mathcal{V} = \{(x_i, y_i) : i \in \{1, \dots, m\} \setminus I\}$

Train the method (i.e. solve problem (8)) using only sample $\mathcal{T}$ as the training data, to obtain $\hat{w}$

Compute test error of $\hat{w}$ on sample $\mathcal{V}$

Repeat k times and average the test errors; the result is an estimate of the *generalization error* using this γ

Repeat for different γ and select the γ that minimizes the estimate

- The above procedure can be costly in time

Ridge Regression Revisited

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m (y_i - w^\top x_i)^2 + \gamma \|w\|^2 \quad (9)$$

- Loss function is the *square loss*, $E(a, b) = (a - b)^2$
- Regularizer is the square of the L_2 norm
- As $\gamma \rightarrow +\infty$, the solution $\hat{w} \rightarrow 0$
- As $\gamma \rightarrow 0^+$, the solution $\hat{w} \rightarrow (X^\top)^+ y$

Note : we always consider $\gamma > 0$; if γ were zero, (9) might have multiple solutions, whereas Problem (9) has a *unique* solution

Minimal Norm Interpolation

- Assume that the linear system

$$X^\top w = y$$

has a solution.

- Then $w_0 = (X^\top)^+ y$ is the solution of the problem

$$\begin{aligned} & \min_{w \in \mathbb{R}^d} \|w\| \\ & \text{s.t. } w^\top x_i = y_i, \quad i = 1, \dots, m \end{aligned} \tag{10}$$

- That is, as $\gamma \rightarrow 0^+$, problem (9) approaches problem (10)

SVM Revisited

The soft margin SVM formulation **(P)** is equivalent to the problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m \max\{1 - y_i(w^\top x_i), 0\} + \gamma \|w\|^2$$

with $\gamma = \frac{1}{2C}$

- The loss function is $E(a, b) = \max\{1 - ab, 0\}$, where $b \in \{-1, 1\}$
- The regularizer is the square of the L_2 norm

SVM Revisited (contd.)

Indeed, we have

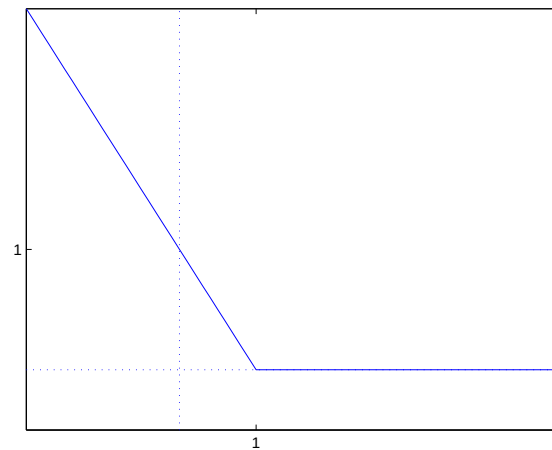
$$\begin{aligned} \min_{w, \xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(w^\top x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned} \quad \Longleftrightarrow$$

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|^2 + C \min_{\xi} \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \xi_i \geq 1 - y_i(w^\top x_i), \quad \xi_i \geq 0 \end{aligned} \quad \Longleftrightarrow$$

$$\min_w \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \max\{1 - y_i(w^\top x_i), 0\}$$

Hinge Loss

The loss function $E(a, b) = \max\{1 - ab, 0\}$ is called the *hinge loss*



Plot of function $\max\{1 - x, 0\}$

No penalty on the “good” side of the margin, i.e. when $\text{sgn}(a) = b$ and $|a| \geq 1$; linear penalty on the “bad” side

Next Lecture

- What is the ridge regression dual?
- What can we say more generally about regularization?
- How do we learn nonlinear functions using kernels?

Bibliography

- SVM: Originally proposed by B. Boser, I. Guyon, and V. Vapnik. In 5th Annual Workshop on Computational Learning Theory, pp. 144–152, ACM, 1992 and extended to the non-separable case by C. Cortes and V. Vapnik. Support-Vector Networks. Machine Learning, 20, 1995.
- See also: Shawe-Taylor and Cristianini's book, Chapter 7.2.