CS 524: Introduction to Optimization
Lecture 30 : Regularization/Classification

Michael Ferris

Computer Sciences Department
University of Wisconsin-Madison

November 13, 2023

Recall: modern regression techniques

Over the last few lectures we've investigated modern regression techniques. Note that linear regression generally has low bias (zero bias, when the true model is linear) but high variance, leading to poor predictions. Modern methods introduce some bias but significantly reduce the variance, leading to better predictive accuracy.
Given a response $y \in \mathbb{R}^m$ and predictors $A \in \mathbb{R}^{m \times n}$, we can think of these modern methods as constrained least squares:

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 \text{ subject to } \|x\|_q \leq t$$

- $q = 1$ gives the lasso (and $\|x\|_1 = \sum_{j=1}^n |x_j|$)
- $q = 2$ gives ridge regression (and $\|x\|_2 = \sqrt{\sum_{j=1}^n x_j^2}$)

See 30tradeoff.ipynb

Often wish to find "sparse" solutions to least-squares problems, in which most of the elements of the solution x are zero. We are willing to sacrifice a little goodness of fit (the objective) in order to find an approximate solution with a small number of nonzeros. The nonzeros are called "explanatory variables;" we wish to find the set of explanatory variables of a given size that best explains the observations.
Talk about ℓ_0 semi-norm. Could do this explicitly by allowing only a certain number of x components to be nonzero, using binary selector variables s_i, $i = 1, 2, \ldots, n$, with constraints $-Ms_i \leq x_i \leq Ms_i$ and $\sum s_i \leq k$, where k is the maximum number of nonzero components. This is a miqcp.

Regularization at large

Most other modern methods for regression can be expressed as

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 \text{ subject to } R(x) \leq t$$

or equivalently

$$\hat{x} = \arg\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 + \lambda \cdot R(x)$$

The term R is called a penalty or regularizer, and modifying the regression problem in this way is called applying regularization

- Regularization can be applied beyond regression: e.g., it can be applied to classification, clustering, principal component analysis
- Regularization goes beyond sparsity: e.g., design R to induce smoothness or structure, instead of pure sparsity

LASSO

This is L_1 penalized linear regression. Problem is really to take weighted sum of objective and ℓ_0 norm of x but we approximate ℓ_0 by ℓ_1 [**?**]. Also could use LASSO, in which we solve the following constrained linear least squares problem for some value of the upper bound T:

$$\min \frac{1}{2}\|Ax - y\|_2^2 \text{ subject to } \|x\|_1 \leq T.$$

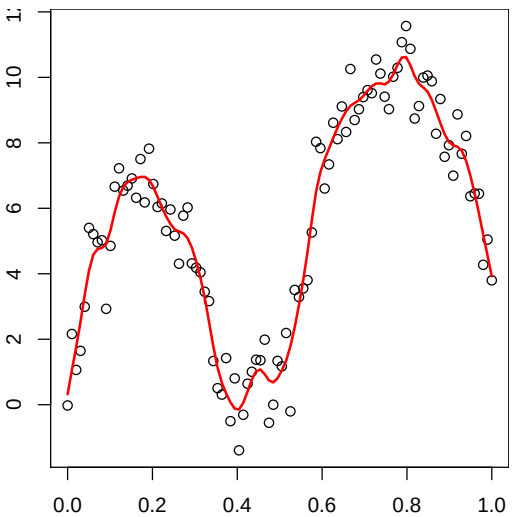
More information at [**?**]. Can formulate as a QP by introducing variables that represent $|x_i|$, $i = 1, 2, \ldots, n$. For small T, the solution contains only a few nonzeros; the number of nonzeros increases as T increases, and for $T = \infty$ we obtain the unconstrained least-squares solution. The idea is to solve for a range of T and observe which components are nonzero, then solve an additional "unconstrained" least squares problem in which just these components are allowed to be nonzero.

Example: smoothing splines

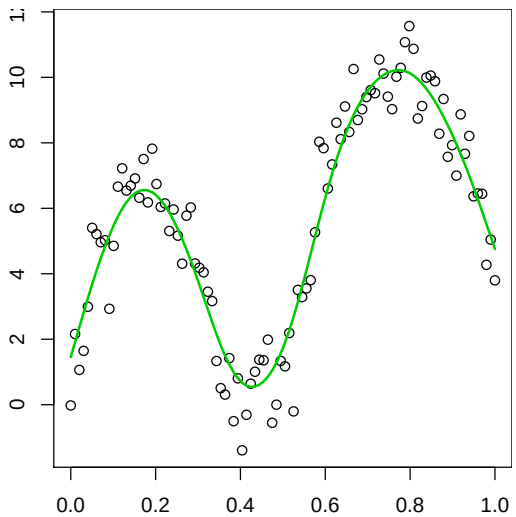
Smoothing splines use a form of regularization:

$$\hat{\phi} = \arg\min_{\phi} \sum_{i=1}^{m} (y_i - \phi(a_i))^2 + \lambda \cdot \underbrace{\int (\phi''(a))^2 da}_{R(\phi)}$$

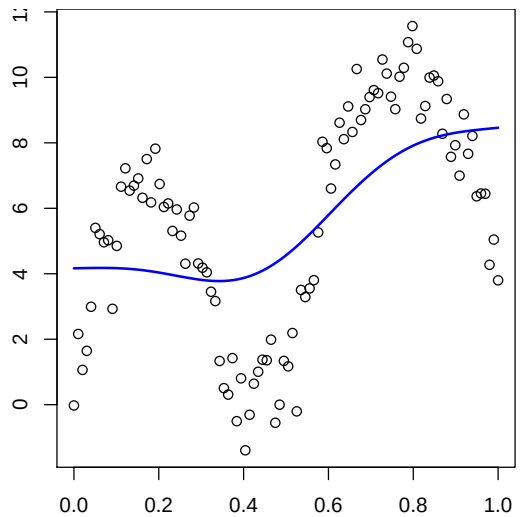
Example with $m = 100$ points:



λ too small



λ just right



λ too big

Classifiers (see `30svm.ipynb`)

Consider now the situation in which we have two sets of points in $\mathbb{R}^n$ which are labeled as P_+ and P_-. Denoting any one of these points by a, we would like to construct a function ϕ so that $\phi(a) > 0$ if $a \in P_+$ and $\phi(a) < 0$ if $a \in P_-$. The function ϕ is known as a *classifier*. Given a new point a, we can use ϕ to classify a as belonging to either P_+ (if $\phi(a) > 0$) or P_- (if $\phi(a) < 0$). An example of such a problem constructs a linear function $\phi(a) = a^T w - \gamma$ to classify fine needle aspirates of tumors as either malignant or benign. We give a brief description of *support vector machines*, a modern tool for classification.

We start by describing the construction of a linear classifier, which has the form $\phi(a) = w^T a - \gamma$, where $w \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$. Ideally, the hyperplane defined by $\phi(a) = 0$ should completely separate the two sets P_+ and P_-, so that $\phi(a) > 0$ for $a \in P_+$ and $\phi(a) < 0$ for $a \in P_-$. If such (w, γ) exist, then by redefining (w, γ) as

$$\frac{(w, \gamma)}{\min_{a \in P_+ \cup P_-} |w^T a - \gamma|},$$

we have that

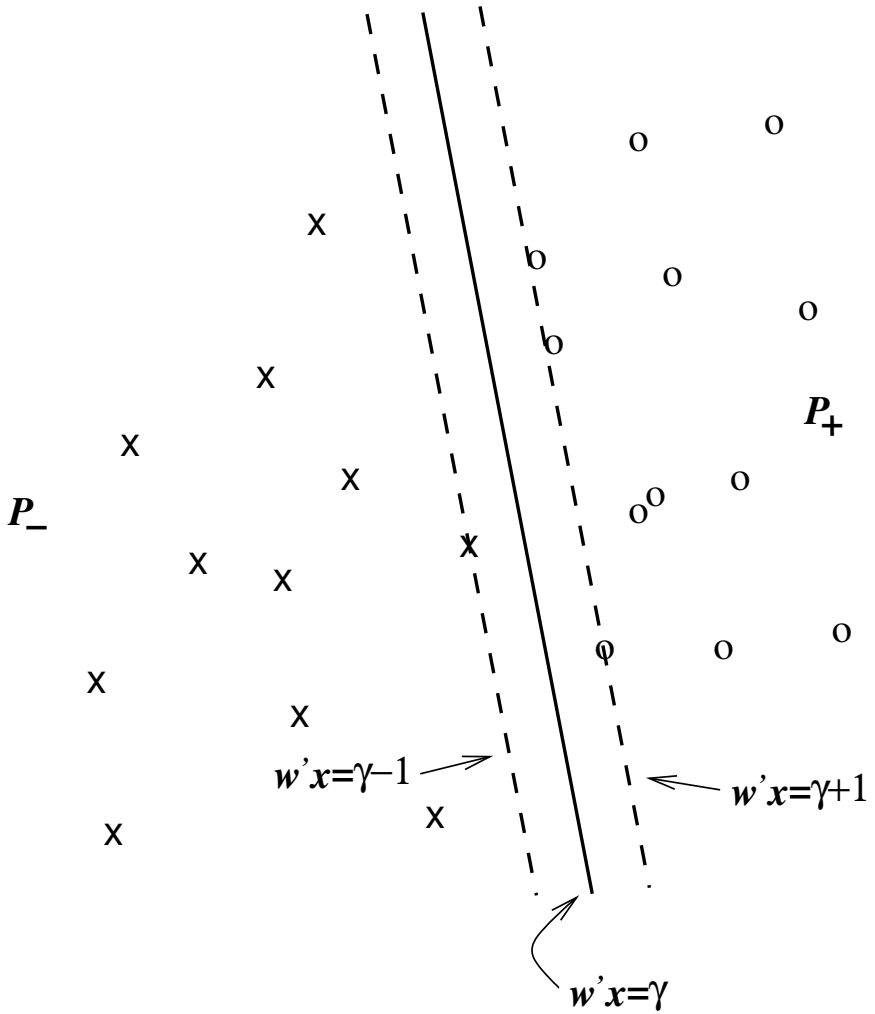
$$\begin{array}{rcl} a \in P_+ & \Rightarrow & \phi(a) = w^T a - \gamma \geq 1, \\ a \in P_- & \Rightarrow & \phi(a) = w^T a - \gamma \leq -1. \end{array} \tag{1}$$

To express these conditions as a system of linear inequalities, we denote the points by a_i and let $m = |P_+| + |P_-|$, where $|P_+|$ and $|P_-|$ denote the number of points in P_+ and P_- respectively. We then define labels y_i for each point a_i as follows:

$$y_i = \begin{cases} 1 & \text{if } a_i \in P_+; \\ -1 & \text{if } a_i \in P_-; \end{cases}$$

The conditions (1) can thus be written succinctly as follows:

$$y_i(a_i^T w - \gamma) \geq 1 \tag{2}$$



shows a separating hyperplane $w^T x = \gamma$ for two point sets, obtained by finding a pair (w, γ) that is feasible for (2) as well as the bounding hyperplanes $w^T x = \gamma \pm 1$.

Margin maximization

If it is possible to separate the two sets of points, then it is desirable to maximize the distance (margin) between the bounding hyperplanes, which is depicted above by the Euclidean distance between the two dotted lines. It can be shown [**?**] that this separation margin is

$$\frac{2}{\|w\|'},$$

where $\|w\|'$ denotes the dual norm. If we take the norm to be the Euclidean (ℓ_2) norm, which is self-dual, then maximization of $2/\|w\|'$ can be achieved by minimization of $\|w\|$ or $\|w\|^2 = w^T w$.

Hence, we can solve the following quadratic program to find the separating hyperplane with maximum (Euclidean) margin:

$$\min_{w,\gamma} \frac{1}{2} w^T w \text{ s.t. } y_i(a_i^T w - \gamma) \geq 1, i = 1, \ldots, m. \tag{3}$$

The *support vectors* are the points a_j that lie on the bounding hyperplanes and such that the corresponding Lagrange multipliers of the constraints of (3) are positive. These correspond to the active constraints in (3).
In practice, it is usually not possible to find a hyperplane that separates the two sets because no such hyperplane exists. In such cases, the quadratic program (3) is infeasible, but we can define other problems that identify separating hyperplanes "as nearly as practicable," in some sense.

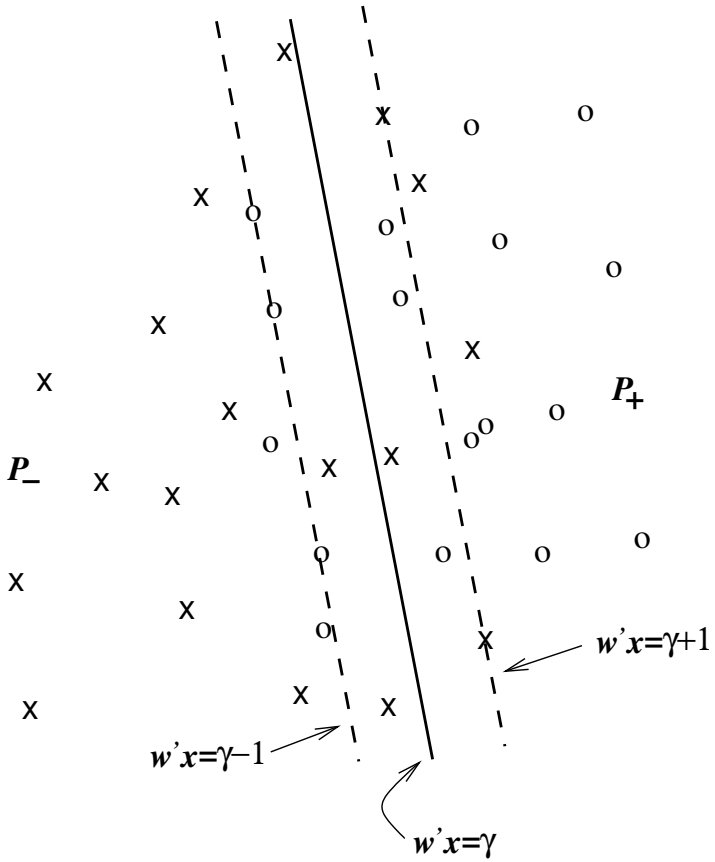
We can define a vector δ whose components indicate the amount by which the constraints (2) are violated, as follows:

$$y_i(a_i^T w - \gamma) + \delta_i \geq 1, \delta_i \geq 0. \tag{4}$$

We could measure the total violation by summing the components of δ, and add some multiple of this quantity to the objective of (3), to obtain

$$\min_{w,\gamma,\delta} \frac{1}{2} w^T w + \nu \sum_i \delta_i \text{ s.t. } y_i(a_i^T w - \gamma) + \delta_i \geq 1, \delta_i \geq 0, \tag{5}$$

where ν is some positive parameter. This problem (5) is referred to as a (linear) *support vector machine* [**?**, **?**, **?**].



shows two linearly nonseparable point sets and the hyperplane obtained by solving a problem of the form (5). The bounding hyperplanes are also shown. In this formulation, the support vectors are the points from each set P_- and P_+ that lie on the wrong side of their respective bounding hyperplanes.

Nonlinear separator

Instead of a linear separator, we can transform the data a_i by some nonlinear transformation ψ and then perform support vector machine classification on the vectors $\psi(a_i)$ instead. The nonlinear *classification function* operates as:

$$\phi(\psi(a)) > 0 \text{ implies } a \in P_+,$$
$$\phi(\psi(a)) < 0 \text{ implies } a \in P_-$$

The separating hypersurface is then $\left\{x : w^T\psi(x) - \gamma = 0\right\}$ and the optimization problem is:

$$\min_{w,\gamma} \nu \sum_{j=1}^{m} \max(1 - y_j(w^T\psi(a_j) - \gamma), 0) + \frac{1}{2}\|w\|_2^2$$

Duality

The dual problem (exercise) is:

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2}\alpha^T Q\alpha - \sum_i \alpha_i \text{ s.t. } 0 \leq \alpha_i \leq \nu, y^T\alpha = 0$$

where

$$Q_{ij} = y_i y_j \psi(a_i)^T \psi(a_j)$$

The solution w of the primal can be recovered as

$$w = \sum_{j=1}^{m} y_j \alpha_j \psi(a_j)$$

and γ is the multiplier on the constraint $y^T\alpha = 0$.
Support vectors are those whose α_i are not zero.

Kernels

The problem can be specified directly using Q instead of via ψ, via a *kernel function* K with $K(a_i, a_j)$ replacing $\psi(a_i)^T\psi(a_j)$. The classifier function then operates as:

$$\sum_{j=1}^{m} y_j\alpha_j\psi(a_j)^T\psi(a) - \gamma = w^T\psi() - \gamma > 0 \text{ implies } a \in P_+,$$

$$\sum_{j=1}^{m} y_j\alpha_j\psi(a_j)^T\psi(a) - \gamma < 0 \text{ implies } a \in P_-$$

which can be evaluated directly via the kernel function as:

$$\sum_{j=1}^{m} y_j\alpha_j K(a_j, a) - \gamma > 0 \text{ implies } a \in P_+,$$

$$\sum_{j=1}^{m} y_j\alpha_j K(a_j, a) - \gamma < 0 \text{ implies } a \in P_-$$

A popular choice of kernel is the Gaussian kernel:

$$K(a_j, a_i) := \exp\left(-\frac{1}{2\sigma}\|a_j - a_i\|^2\right)$$

with related Gram matrix Q given by

$$Q_{ij} = y_i y_j K(a_i, a_j)$$

can be used in the model to determine w (essentially α) and γ for the classifier.
A function $K(x, y)$ is a valid kernel if it corresponds to an inner product in some (perhaps infinite dimensional) feature space. General condition: construct the Gram matrix $K(a_i, a_j)$ check that it's positive semidefinite, (see Mercer's condition).