REGULARIZATION

Regularization is the idea to promote "simple" solutions to machine learning problems. This is typically achieved by adding a "regularization term" to the objective function:

Examples

Least squares linear regression:

$$\sum_{i=1}^{n} (y_i - \beta^T x_i - \beta_0)^2 + \lambda \|\beta\|^2$$

Logistic regression

$$-\log \ell(\theta) + \lambda \|\theta\|^2$$

Here $\lambda > 0$ is a "tuning parameter" that controls the tradeoff between data fit and complexity.

Why regularize?

Well-conditioned Hessians

In high-dimensional problems, the Hessian can be singular or ill-conditioned.

Examples

1. In least squares linear regression, the Hessian of the sum of squared errors is A^TA. If $d > n$, this matrix is singular

2. In logistic regression, as seen on the homework, the Hessian can be ill-conditioned for high-dimensional problems.

We will see that regularization leads to well-conditioned Hessians.

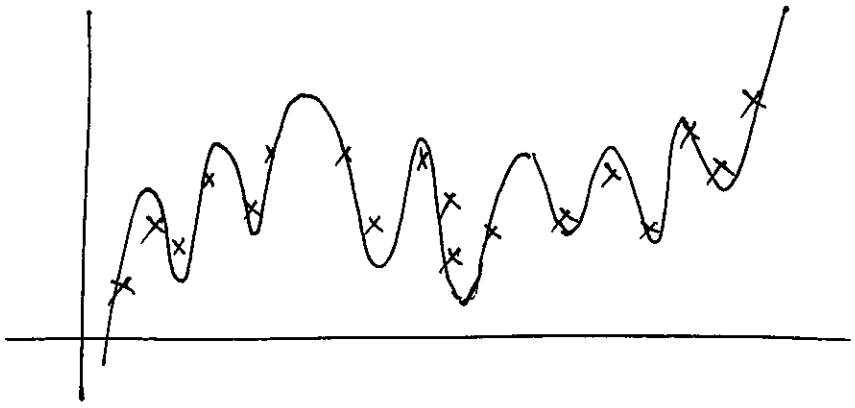
Avoid overfitting

Sometimes "simpler" solutions are better because they are less likely to "overfit."

Example] Suppose we want to use polynomial regression, but we don't know the right degree, so we just choose a very high degree:

$$\Phi(x) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^{20} \end{bmatrix}^T$$

We are very likely to overfit:



By regularizing, we may achieve a smoother fit that will generalize well to new examples.

In fact, these two motives for regularization are two sides of the same coin:

- If the Hessian is ill-conditioned
  it has an eigenvalue very close to zero.
  In the direction of the corresponding
  eigenvector, the objective function is very "flat."
  Along this direction, all solutions are
  equally good in terms of the objective function,
  but many of those solutions will overfit.

- In LS polynomial regression, the Hessian
  quickly becomes ill-conditioned as the degree
  increases.

Ridge Regression

Given $y_i = f(x_i) + \varepsilon_i$, where $f(x_i) = \beta^T x_i + \beta_0$.

Instead of minimizing the sum of squared errors, in ridge regression, we minimize

$$\sum_{i=1}^{n} (y_i - \beta^T x_i - \beta_0)^2 + \lambda \|\beta\|^2$$

where $\lambda > 0$ is a tuning parameter.

Note: β_0 is not penalized so that our solution is independent of where the origin is located

Let's derive the solution. First, let's eliminate β_0

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta^T x_i - \beta_0) = 0$$

$$\Rightarrow \hat{\beta}_0 = \frac{1}{n} \sum_i y_i - \hat{\beta}^T x_i$$

$$=$$

(A)

Thus, we are left to minimize ...

$$\sum_{i=1}^{n} \left(y_i - \bar{y} - \beta^T (x_i - \bar{x}) \right)^2 + \lambda \beta^T \beta$$

wrt β. For convenience, let $\tilde{x}_i = x_i - \bar{x}$, $\tilde{y}_i = y_i - \bar{y}$

The criterion may be written

$$(\tilde{y} - A\beta)^T (\tilde{y} - A\beta) + \lambda \beta^T \beta, \quad A = \begin{bmatrix} \tilde{x}_1^{(1)} & \cdots & \tilde{x}_1^{(d)} \\ \vdots & & \\ \tilde{x}_n^{(1)} & \cdots & \tilde{x}_n^{(d)} \end{bmatrix}$$

$$\tilde{y} = [\tilde{y}_1 \cdots \tilde{y}_n]^T$$

$$= \tilde{y}^T \tilde{y} + \beta^T A^T A \beta - 2\beta^T A^T \tilde{y} + \lambda \beta^T \beta$$

$$= \beta^T \left[A^T A + \lambda I \right] \beta - 2\beta^T A^T \tilde{y} + \tilde{y}^T \tilde{y}$$

$$\frac{\partial}{\partial \beta} = 0 \implies (A^T A + \lambda I)\beta = A^T \tilde{y}$$

(B) $\implies$

Observations
- $\lambda = 0$ recovers least-squares linear regr.
- λI increases the eigenvalues of $A^T A$ by λ, so that $A^T A + \lambda I$ is not ill-conditioned.

Soft Margin Hyperplane

The training error of a linear classifier may be bounded as

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i(w^T x_i + b) < 0\}}$$

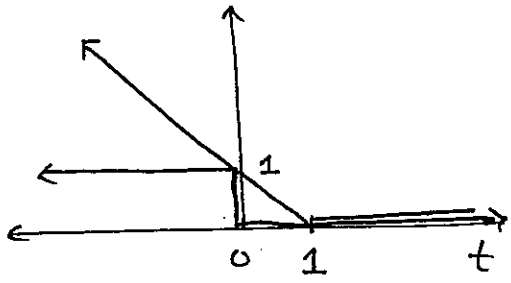
$$\leq \frac{1}{n}\sum \phi\left(y_i(w^T x_i + b)\right)$$

where $\phi(t)$ is any upper bound on $1_{\{t<0\}}$.

Let's take

$$\phi(t) = \max\{0, 1-t\}$$

$$=: (1-t)_+$$



In addition, let's add a quadratic penalty to keep the coefficients small.

$$\Longrightarrow \min_{w,b} \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\left(1 - y_i(w^T x_i + b)\right)_+$$

Compare this to the quadratic program for the optimal soft-margin hyperplane:

$$\min_{w,b,\xi_i} \quad \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1,\dots,n$$

$$\xi_i \geq 0, \quad i = 1,\dots,n$$

Claim If $C = \frac{1}{\lambda}$, these two optimization problems are solved by the same w, b.

© Proof:

Key

A $\bar{y} - \hat{\beta}^T \bar{x}$ B $\hat{\beta} = (A^TA + \lambda I)^{-1} A^T \underline{y}$

C Since scaling an objective function by a positive constant does not change the solution, it suffices to show that

OP1 $\min_{(w,b)} \frac{1}{2}\|w\|^2 + \frac{1}{n\lambda} \sum (1 - y_i(w^T x_i + b))_+$

OP2 $\min_{(w,b,\xi)} \frac{1}{2}\|w\| + \frac{1}{n\lambda} \sum \xi_i$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$

are solved by the same (w, b).

- Suppose (w^*, b^*) is an optimizer of OP1

Claim: (w^*, b^*, ξ^*) is an optimizer of OP2, where

$$\xi_i^* = \max(0, 1 - y_i(w^{*T} x_i + b^*))$$

Suppose not, and let (w, b, ξ) be an optimizer of OP2.
Since (w, b, ξ) is a global optimizer

- If $\xi_i > 0$, then $y_i(w^T x_i + b) = 1 - \xi_i$
  [otherwise, we could decrease the objective function without violating the constraints]

- If $\xi_i = 0$, then $y_i(w^T x_i + b) \geq 1$

Thus $\sum \xi_i = \sum (1 - y_i(w^T x_i + b))_+$ and so

$$\frac{1}{2}\|w\|^2 + \frac{1}{n\lambda} \sum (1 - y_i(w^T x_i + b))_+$$

$$= \frac{1}{2}\|w\|^2 + \frac{1}{n\lambda} \sum \xi_i$$

$$< \frac{1}{2}\|w^*\|^2 + \frac{1}{n\lambda} \sum \xi_i^*$$

$$= \frac{1}{2}\|w^*\| + \frac{1}{n\lambda} \sum (1 - y_i(w^{*T} x_i + b))_+$$

Which contradicts optimality of (w^*, b^*) for OP1.

- Suppose (w^*, b^*, ξ^*) is an optimizer of OP2.
  Claim: (w^*, b^*) is an optimizer of OP1.
  The argument is similar and is left as an exercise.