Jun 9, 2014

l_0, l_1, l_2 Regularization

The L_q-Norms

Given a vector $\mathbf{x} = (x_1, \ldots, x_n)$ we define the l_q-norm for $n > 0$ as follows:

$$\|\mathbf{x}\|_q = \sqrt[q]{\sum |x_i|^q}$$

In the case where $q = 0$ we define l_0-norm to be the number of non-zero elements of the vector:

$$\|\mathbf{x}\|_0 = \#(i|x_i \neq 0)$$

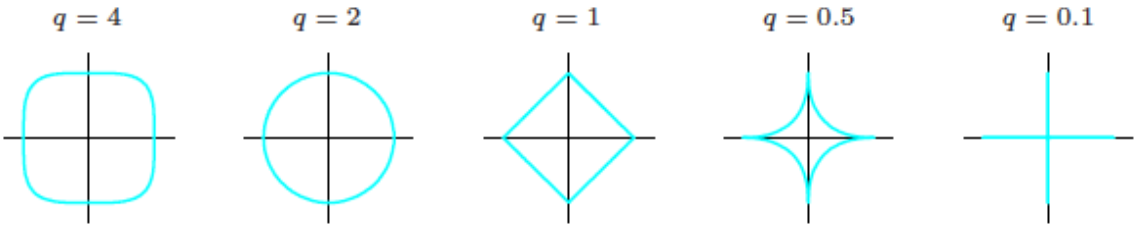
For a data set $X = \{x_1, \ldots, x_n\}$, three measures of central tendency, mode, median and mean, can be defined in terms of the l_0, l_1, and l_2 norms respectively. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (1, \ldots, 1)$ be $1 \times n$ vectors:

$$\text{mode of } X = \underset{y}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|_0$$

$$\text{median of } X = \underset{y}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|_1$$

$$\text{mean of } X = \underset{y}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|_2$$

In two dimensions, if we plot the contours of $z = \|(x_1, x_2)\|_q^q = |x_1|^q + |x_2|^q$ we get:



(Image from C.M. Bishop, *Pattern Recognition and Machine Learning*; Springer, 2006)

Note that $q = 1$ is the smallest value of q for which the contour plots define convex regions.

l_2 Regularization

In order to address overfitting, we can use l_q regularizers to shrink model parameters by imposing a penalty on their size. Suppose $K(\theta)$ is the objective function for a model with parameters θ_i for $i = 1, \ldots, d$ that would like to minimize on a training set. We add the square of the l_2-norm of the parameter vector to $K(\theta)$ to obtain the objective function

$$J(\theta) = K(\theta) + \lambda \|\theta\|_2^2 = K(\theta) + \lambda \sum_{i=1}^{d} \theta_i^2, \ \ \lambda > 0$$

In the case of linear regression we want to minimize

$$J(\theta) = (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \lambda \theta^T \theta.$$

This case, called ridge regression, results in the solution

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^T \mathbf{y}$$

a natural hack to address problems that arise from a poorly conditioned system of equations.

Obviously, the larger the magnitude of λ the larger the penalty on the parameters and as $\lambda \to \infty$ we get that $\theta_i \to 0$ for $i = 1, \ldots d$. If we interpret the minimization of $J(\theta)$ as a constrained optimization problem

$$\min K(\theta) \text{ subject to } \sum_{i=1}^{d} \theta_i^2 \leq t(\lambda),$$

we see that the values $\lambda \in [0, \infty)$ generate a path of solutions in $\mathbb{R}^d$ from the unconstrained solution of the minimization ($\lambda = 0$) to $\theta = 0$ ($\lambda \to \infty$). It is important to note for comparison with other regularizers that solutions occur at tangency points of level curves of the functions $\|\theta\|_2^2$ and $J(\theta)$.

l_1 Regularization

If we impose a penalty on the objective function using the l_1-norm, we get

$$J(\theta) = K(\theta) + \lambda \|\theta\|_1 = K(\theta) + \lambda \sum_{i=1}^{d} |\theta_i|, \;\; \lambda > 0$$

This method is called the LASSO (least absolute selection and shrinkage operator).

This penalty generally induces more sparse solutions than the l_2-norm penalty as values of θ tend to shrink much faster as λ grows. If we examine the level curves of the constrained interpretation as above, we see that, for nicely behaved objective functions, it's much more likely for level curves of $J(\theta)$ to be tangent to level curves of the l_1-norm at its corners where some elements of θ are identically zero. However, since the l_1-norm is not a differentiable function at the origin, optimization is often not as straight forward as with the l_2 regularization.

Note that a linear combination of the l_1 and l_2 penalties is called elastic net regularization.

l_0 Regularization

If we impose a penalty on the objective function using the l_0-norm, we get

$$J(\theta) = K(\theta) + \lambda \|\theta\|_0 = K(\theta) + \lambda \#(i|\theta_i \neq 0), \;\; \lambda > 0$$

which penalizes the objective function by a multiple of the number of non-zero parameters and is called best subset selection. If we consider the constrained optimization problem

$$\min K(\theta) \text{ subject to } \#(i|\theta_i \neq 0) \leq n,$$

we see that the solution is forced to be at a point where a level curve of $K(\theta)$ intersects at least $d - n$ coordinate axis. Although a natural regularizer to define, the l_0-norm is not convex so this optimization is often computationally difficult to solve exactly. The l_0-optimization problem is known to be NP-hard.

l_0, l_1, l_2 Regulerizers from Bayes Estimates

l_0, l_1, and l_2 regularization can all be derived from Bayes estimates with different priors. l_2 regularization is equivalent to using a circular Gaussian conjugate prior with $\theta_0 = 0$ and variance $\Sigma = \tau I_d$. l_1 regularization is equivalent to using a Laplace prior with mean $\theta_0 = 0$. While l_2 results from the posterior mean, since the posterior is Gaussian, the other two regularizers result from the modes of the posterior. In the case of l_1 regularization, another way to account for the sparse solutions is the large "spike" mode of the Laplace prior at zero. However, since Laplace distributions are more fat-tailed than Gaussians, Bayesian inference (marginalization) or other non-mode estimates will not result in sparse use of the features.