Modern Computational Statistics lec02 Notes
Part I

This is based on PKU course: Modern Computational Statistics. Thanks to Prof. Cheng Zhang, this is a very interesting course.

This lecture will introduce some of convex optimization. You can also see CMU's course convex optimization by Ryan Tibshirani, for more details.

1 Normal Optimization Methods

1.1 Least Square Regression Models

Common least square regression model can be expressed as,

$$minimize L(\beta) = \frac{1}{2}||Y - X\beta||^2$$

This is a quadratic problem. We can solve it by setting the gradient to zero.

$$\nabla_\beta L(\beta) = -X^T(Y - X\hat{\beta}) = 0$$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

given that the Hessian is positive definite,

$$\nabla^2 L(\beta) = X^TX > 0$$

1.2 Regularized Regression Models

The common regularized regression models we know are ridge regression and lasso."Regularization" is a way to give a penalty to certain models[1]. Ridge regression belongs to L_2 penalty, which gives penalty to the $l_2 - norm$. Because all coefficients are shrunk by the same factor, so all the coefficients remain in the model. And lasso uses the $l_1 - norm$ for penalty, which limits the size of the coefficents. This sometimes results in the eliminationof some coefficents, making sparse models.

Below is the common approach of using regression models:

$$minimize \quad L(\beta) = \frac{1}{2}||Y - X\beta||^2$$

$$subject \quad to \sum_{j=1}^{p} |\beta_j|^\gamma \leq s$$

This regularized regression expression is called the Bridge regression model. When $\gamma = 1$, it represents the lasso regression, and when $\gamma = 2$, it represents the ridge regression. When bridge regression allows for $\gamma < 1$, it gives a

[1] https://www.statisticshowto.datasciencecentral.com/regularized-regression/

nonconvex regression. It can be used to select groups of covariates,especially from sparse data. See more for Bridge penalty vs. Elastic Net regularization on https://stats.stackexchange.com/questions/224531/bridge-penalty-vs-elastic-net-regularization.

1.3 General Optimization Problems

Below gives the general for of optimization problems:

$$minimize \quad f_0(x)$$

$$subject \quad to \quad f_i(x) \leq 0, i = 1, ..., m$$

$$h_j(x) = 0, \quad j = 1, ..., p$$

If the objectiven function $f_0(x)$, the inequality constraints $f_i(x)$ and the equality constraints $h_j(x)$ are all convex function, this problem is a convex optimization problem.

2 Convex Optimization Problems

2.1 Convex Sets

Convex set: $C \subseteq R^n$ such that

$$x, y \in C \quad \rightarrow \quad tx + (1-t)y \in C \quad for \quad all \quad 0 \leq t \leq 1$$

To discribe it in a more common way, if we have two dots in a set, draw a line to connect the dots, then this line should be contained in the set. Like:

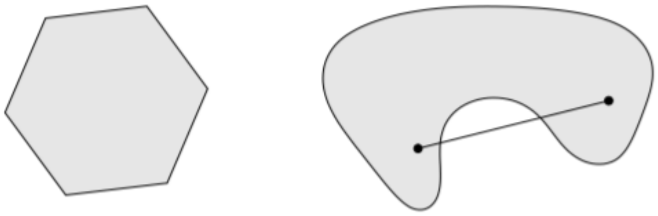

Figure 1: Convex set

The left one is convex set but the right one is not.
Some propositions:

- The joint of some convex sets is still convex set.

- The linear combination of convex sets is called a convex combination.
- The closure and interior of a convex set are convex sets.

[1] See Convex Optimization Theory for proofs.

2.2 Convex Functions

Definition: A function f: $R^n \rightarrow R$ is convex if its domain D_f is a convex set,and $forall x, y \in D_f$ and $0 \leq \theta \leq 1$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

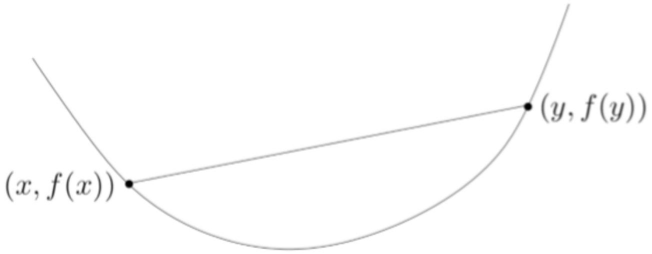

Figure 2: Convex Function

We can see from the figure that the connection of the two points is always larger than the actual value. For example, affine function and norms are always convex functions.

Firt-order Characterization: [2] If f is differentiable, then f is convex iff $dom(f)$ is convex, and

$$f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

for all $x, y \in dom(f)$. Therefore for a differentiable convex function

$$\nabla f(x) = 0 \quad \Leftrightarrow \quad x \quad minimizes \quad f$$

Firt-order Characterization: [2] If f is twice differentiable, then f is convex iff $dom(f)$ is convex, and $\nabla^2 f(x) \succeq 0$ for all $x \in dom(f)$.

Jensen's Inequality: [2] If f is convex, and X is a random variable supported on $dom(f)$, then

$$f(E(X)) \leq E(f(X))$$

.

2.3 Basic Terminology and Notations

If x is feasible and $f(x) = f^*$, then x is called optimal. If x is feasible and $f(x) \leq f^* + \epsilon$, then x is called $\epsilon - suboptimal$.

Let X_{opt} be the set of all solutions of convex problem, then

$$X_{opt} = argmin \quad f(x)$$

$$subject \quad to \quad g_i(x) \leq 0, i = 1, ..., m$$

$$Ax = b$$

Here X_{opt} should be convex set. And if f is strictly convex, then the solution is unique.

2.3.1 Optimality Criterion

Assuming f_0 is convex and differentiable, x is optimal iff

$$\nabla f_0(x)^T(y - x) \geq 0, \forall feasible y$$

And for unconstrained problems, x is optimal iff

$$\nabla f_0(x) = 0$$

2.3.2 Local Optimality

x is locally optimal if for a given $R > 0$, it is optimal for

$$minimize \quad f_0(z)$$

$$subject \quad to \quad f_i(z) \leq 0, i = 1, ...m$$

$$h_j(z) = 0, j = 1, ..., p$$

$$||z - x|| \leq R$$

In Tibshirani's notes, this is expressed as:

$$f(x) \leq f(y)$$

for all feasible y such that $||x - y||_2 \leq R$ **For convex optimization problems, local optima are global optima.**

2.3.3 Examples

Consider LASSO and SVM, do they have unique solution?

2.4 The Lagrangian

Consider a general optimization problem

$$minimize \quad f_0(x)$$

$$subject \quad to \quad f_i(x) \leq 0, i = 1, ..., m$$

$$h_j(x) = 0, j = 1, ..., p$$

Define the Lagrangian $L :^n \times R^m \times R^p \to R$ as

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x)$$

where λ and v are dual variables or $Lagrangian \quad multipliers$.

The **Lagarangian dual function** is defined as

$$g(\lambda, v) = \inf_{x \in D} L(x, \lambda, v)$$

The dual function is the pointwise infimum of a family of affine functions of (λ, v), it is concave, even when the original problem is not convex.

If $\lambda \geq 0$, for each feasible point x,

$$g(\lambda, v) = \inf_{x \in D} L(x, \lambda, v) \leq L(x, \lambda, v) \leq f_0(x)$$

Therefore, $g(\lambda, v)$ is a lower bound for the optimal value.

$$g(\lambda, v) \leq p^*, \forall \lambda \geq, v \in R^p$$

And to find the best lower bound, we should solve the Lagrangian dual problem.

$$maximize \quad g(\lambda, v), \quad subject \quad to \quad \lambda \geq 0$$

This is a convex optimization problem. Solving the dual problem and getting the corresponding solution (λ^*, v^*) can solve the primal problem.

2.5 Weak and Strong Duality

Weak Duality: If dual optimal value is g^*,then $f^* \geq g^*$.(This always holds even if primal problem is nonconvex).

Strong Duality: When f^* = g^*.

Slater's condition: If the primal is a convex problem,and there exists at least one strictly feasible $x \in R^n$,meaning

$$h_1(x) < 0, ..., h_m(x) < 0 \quad and \quad l_1(x) = 0, ..., l_r(x) = 0$$

then Strong Duality holds.

References

[1] Convex Optimization Theory, by Dimitri P.Bertsekas.

[2] Lecture Notes for Convex Optimization, by Ryan Tibshirani,lec2, p21.