Commons (4/12/2018)

It is always good to develop intuitions and pictures in learning. If one wants to remember relations between “variance-bias trade-off”, one could imagine driving a car: to keep variance down by not turning the tearing wheel much but resulting a big bias or to keeping turning left and right to minimize bias as the expense of increasing variance…machine (brain) is often to keep a fair balance!

***Common1***: **Lagrange multiple**



A mechanics toy model: a string has a relaxation length L0, without any constraint (A in Fig), the energy is

f(x) = (x – L0)^2 /2 convex function of x

Min\_x f(x) = 0 because d f(x)/d x = (x \*– L0) =0, at min f, x\*=L0; f\*(L0)=0.

If there is a constraint, x <= L1 < L0 (B in Fig), or g(x) = x – L1 = 0, using Langrage multiplier λ >= 0 to turn this “Primal” optimization problem into a “Dual” optimization problem **without a constraint**:

L(x, λ) = E(x) + λ g(x) = (x – L0)^2 /2 + λ (x – L1)

Dual step one: min\_x L(x, λ)

d L(x, λ) /dx = (x – L0) + λ = 0 hence at min L, x\*= L0 – λ;

L\*( λ) = E(x\*( λ)) + λ g(x\*( λ)) = (x\* -L0)^/2 + λ (x\* - L1) = λ^2/2 + λ(L0 - λ -L1]= - λ^2/2 + λ(L0 –L1)

= - [λ –(L0-L1)]^2/2 + (L0 – L1)^2/2 now a concave function!

Dual step two:

Max\_ λ L\*( λ) = (L0-L1)^2/2 (Spring is compressed, by Hook’s law, energy is increased to this)

because dL\*( λ)/d(λ) = - [λ\* – (L0 – L1)] = 0, λ\* = L0-L1 (we can see the physical meaning of λ\* is the “virtual force” that the constraint must exert to the ball in order to balance it at the new equilibrium position L1 < L0). If there were another constraint L2 < L1, one can introduce two Langrage multipliers and convince oneself, the “virtual force” only needs to balance L2 (“support vector”) which would automatically take care of the constraint at L1 for which the “virtual force” λ would be zero (“sparsity”)!

Comment2: **Geometry of Primal-Dual in Convex Optimization**

**Primal**: inf f(x) with constraint g(x) <= 0 (one can change the sign if constraint is >=0)

Suppose point x in a convex space X is mapped into a point (g(x),f(x)) in a convex space G

 

Then, the primal problem consists in finding a point in G with y ≤ 0 that has minimum ordinate z. Obviously this point is (y¯, z¯).

**Dual**: Max θ(u), subject to: u ≥ 0, θ(u) = inf {f (x)+u g(x) : x ∈ X}.

Given u ≥ 0, the Lagrangian dual subproblem is equivalent to minimise z + uy over points (y, z) in G. Note that z + uy = α is the equation of a straight line with slope −u that intercepts the z-axis at α

 

In order to minimise z + uy over G we need to move the line z + uy = α parallel to itself as far down as possible, whilst it remains in contact with G. The last intercept on the z-axis thus obtained is the value of θ(u) corresponding to the given u ≥ 0. Finally, to solve the dual problem, we have to find the line with slope −u (u ≥ 0) such that the last intercept on the z-axis, θ(u), is maximal. Such a line has slope −u¯ and supports the set G at the point ( ¯ y, ¯ z). Thus, the solution to the dual problem is ¯ u, and the optimal dual objective value is ¯ z.

The solution of the Primal problem is ¯ z, and the solution of the Dual problem is also ¯ z. It can be seen that, in the example illustrated, the optimal primal and dual objective values are equal. In such cases, it is said that there is no duality gap (strong duality).

***Comment3* Kernel K(x,y) >0, symmetric** 

This condition means, the geometry is a Elliptic curve. For example, when K is 2X2 matrix, a vector <V|=(v1,v2), quadratic equation <V|K|V> = K11\*v1^2 + 2K12\*v1\*v2 + K22\*v2^2=R^2 (constant) defines the ellipse. One could idealize K by orthogonal matrix O (rotation) of the coordinates (see fig), so that O|V> = |V’>, <V|K|V> = <V|OTDO|V> = <W|D|W>=R^2, the diagonal matrix has two positive eigenvalues Diag(D)=(d1,d2)=(a, b). If a=b, ellipse becomes a circle, the quadratic equation becomes w1^2+w2^2 =r^2 where r^2=R^2/a. Since K(x,y)= d1\*e1(x) e1(y)+ d2\*e2(x) e2(y), e1,e2 are the two eigenvectors. If K(x,y)= <φ(x)| φ(y)>, we do not need to know )| φ(x)> explicitly; even if it is infinitely dimensional, the effective dimension is number of parameters in the Kernel!