Machine Learning Sheet 02

3) Mean Shift

a) $x_j^{t+1} = x_j^t + \alpha_j^t \frac{2}{n} \sum_{i:\|x_i - x_j^t\| < 1} (x_i - x_j^t)$

Goal: update to be equivalent to moving to the local mean of points within the kernel support

Idea: set α_j^t so that update vector moves x_j^t directly to the mean of nearby points.

$N_j = \sum_{i:\|x_i - x_j^t\| < 1} 1$; local mean $= \frac{1}{N_j} \sum_{i:\|x_i - x_j^t\| < 1} x_i$

$\Leftrightarrow$ local mean $- x_t^j = \frac{1}{N_j} \sum_{i:\|x_i - x_j^t\| < 1} (x_i - x_j^t)$

$x_{t+1}^j = x_t^j + \alpha_j^t \frac{2}{n} N_j \left(\frac{1}{N_j} \sum_{i:\|x_i - x_j^t\| < 1} (x_i - x_j^t) \right)$

$= x_j^t + \alpha_j^t \cdot \frac{2N_j}{n}$ (local mean $- x_t^j$).

$\Rightarrow \alpha_j^t = \frac{n}{2N_j}$

sensible because:

- learning rate adapts to the local density of points
  many nearby points $\rightarrow$ small stepsize and vice versa

- ensures that x_j^t is moved directly to the mean of nearby points $\Rightarrow$ convergence to local modes.

5) Linear Regression: Heteroscedastic Noise

$y_n = \beta^T x_n + \varepsilon_n \quad E[\varepsilon_n] = 0 \quad Var[\varepsilon_n] = \sigma_n^2$

minimize the weighted sum of squares of the residuals

$J(\beta) = \sum_{n=1}^{N} (y_n - \beta^T x_n)^2 \longrightarrow \sum_{n=1}^{N} (y_n - \beta^T x_n)^2 / \sigma_n^2$

observations with lower variance contribute more to the determination of the coefficients β that the ones with higher variance. By minimizing $J(\beta)$ we find the best estimate.

$\hat{\beta} = (X^T W X)^{-1} X^T W y$ with W diagonal $W_{nn} = \frac{1}{\sigma_n^2}$

y observed values X matrix of predictors

$E[\hat{\beta}] = E[(X^T W X)^{-1} X^T W y]$

$= (X^T W X)^{-1} X^T W E[y] = (X^T W X)^{-1} X^T W (X\beta)$

$= (X^T W X)^{-1} X^T W X \beta = \beta$

$Cov(Ay) = A\, Cov(y) A^T$; $Cov(y) = Cov(X\beta + \varepsilon) = Cov(\varepsilon) = diag(\sigma_1^2, \sigma_2^2, \dots \sigma_N^2)$.

$Cov(y) = a^2 I$

$\Rightarrow Cov(\hat{\beta}) = (X^T W X)^{-1} X^T W\, Cov(y) W X (X^T W X)^{-1}$

$= (X^T W X)^{-1} X^T W W X (X^T W X)^{-1} a^2$

$= (X^T W X)^{-1} X^T W X (X^T W X)^{-1} a^2 = a^2 (X^T W X)^{-1}$

a^2 is a scaling factor, this accounts for the diffront variances.