Name: Mandar Darwatkar

SID: 861141010

Date: November 14, 2014

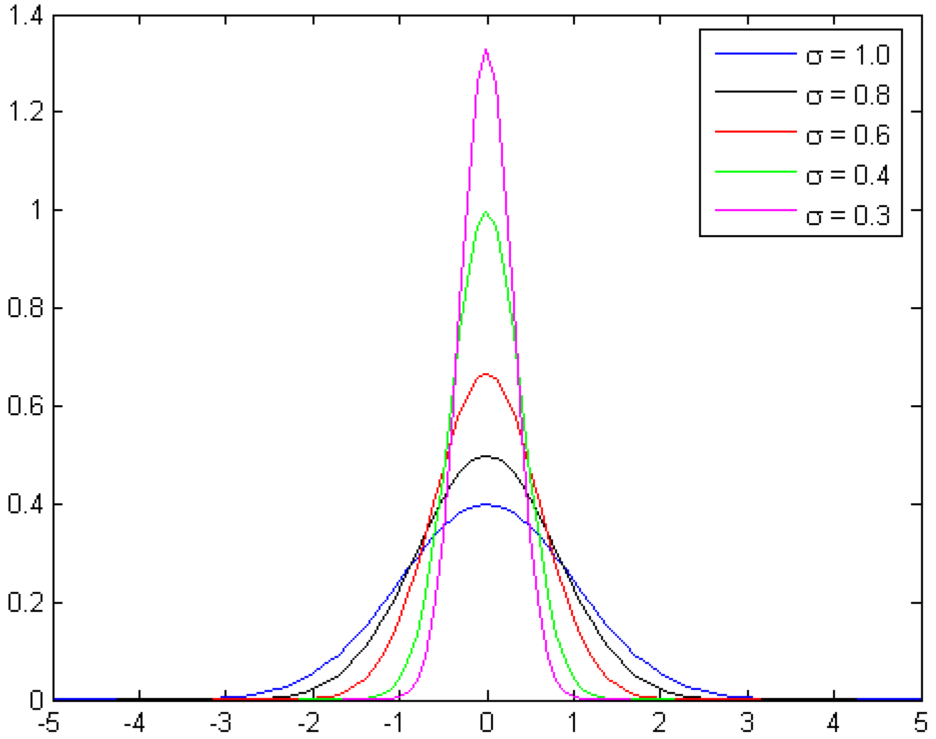
FALL 2014 - CS229

PS5 – part c solution

The kernel width is basically the kernel smoothing parameter. In other words we can say that it determines the width of local neighbourhood.

The larger the width is, kernel considers (averages out on) more observations, hence we get less variance but high bias. Whereas, small kernel width gives high variance but low bias. Therefore, when the data points are linearly inseparable, then the kernel width should be small in order to classify with lesser error (because for large values of width, the classifier becomes almost linear).

When the kernel (Gaussian) is normalized, the area under the curve is always unity. When the width reduces, amplitude increases. In other words, each function is independent of other and the data is classified as if it were memorized i.e. kernel is overfitting. This is because small width implies less number of observations are considered in neighbourhood (the data points close to given center). Please see the illustrations.



$$\lim_{\sigma^2->0} \frac{e^{\frac{-|x-x_i|^2}{2\sigma^2}}}{\sum_j e^{\frac{-|x-x_j|^2}{2\sigma^2}}}$$

$$= \lim_{\sigma^2->0} \frac{e^{\frac{-|x-x_i|^2}{0}}}{\sum_j e^{\frac{-|x-x_j|^2}{0}}}$$

$$= \lim_{\sigma^2->0} \frac{e^{-\infty}}{\sum_j e^{-\infty}}$$

$$= \lim_{\sigma^2->0} \frac{0}{0}$$

= undefined

In other words, model is trained by maximizing its performance on some set of training data. However, its efficacy is determined by its ability to perform well on unseen data. Consequently, the model is going to poorly perform on unseen records because even the minor fluctuations are drastically amplified.