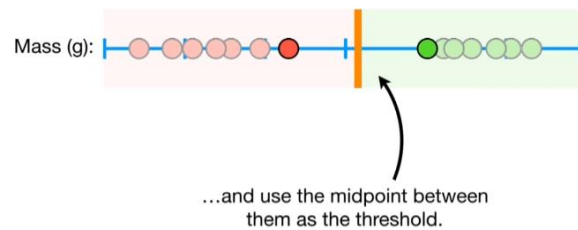
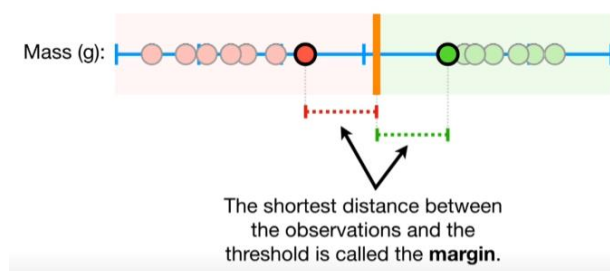


Support Vector Machine SVM

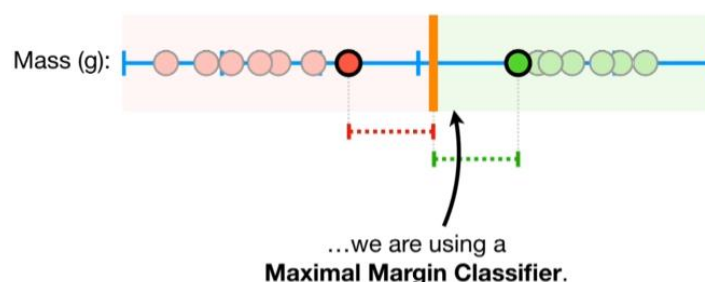
Maximal Margin Classifier



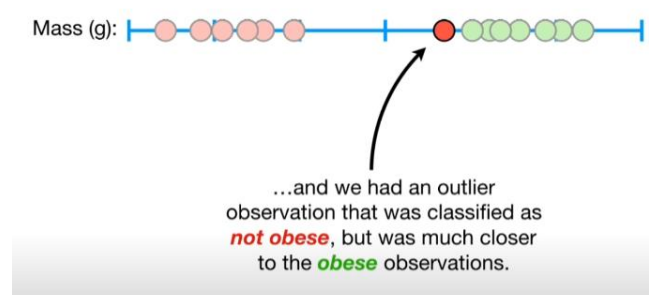
在一般的情况下做分类，会设置一个 Threshold，当这个输入值小过 Threshold 就是一个 Class，当大过 Threshold 就是另一个 Class。如上图设置的 Threshold 是以左边的 Class 最大值与右边的 Class 最小值取平均。



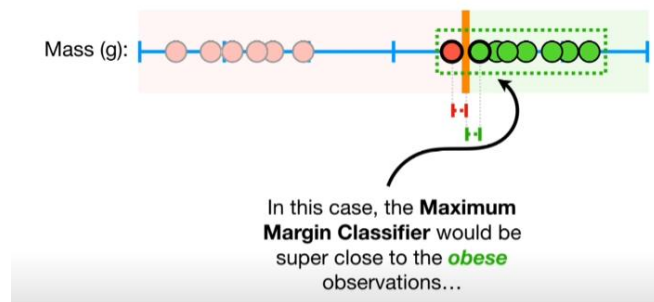
如上图，Data Point 与 Threshold 的距离也被称之为 Margin。



当设置的 Threshold 是在两个 Observations 的中间距离，这时候的 Margin 也是最大的，也被称之为 Maximal Margin Classifier。

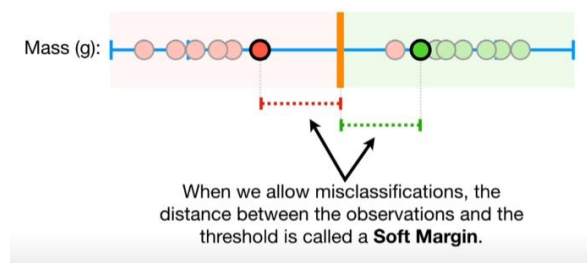


当 Class 1 有一个 Outlier 非常接近 Class 2 的时候，使用 Maximal Margin Classifier 就会有问题。

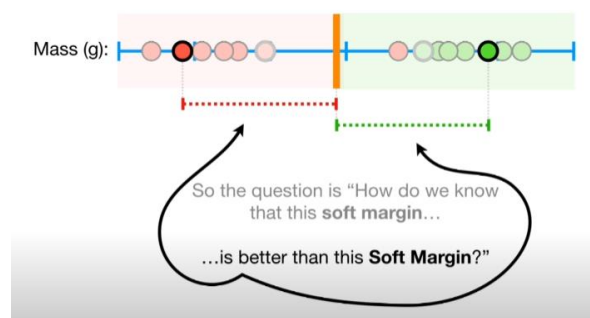


上图分类的是 Class 1 Outlier 与 Class 2 最小值之间的平均距离当作 Threshold。这时候的准确性就不高，High Variance。

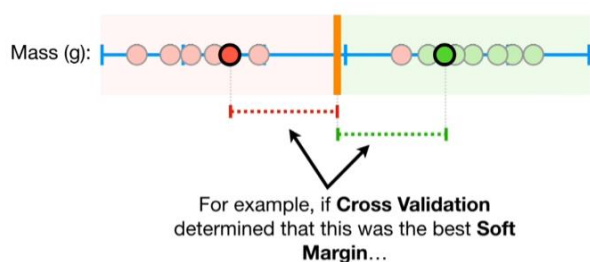
Support Vector Classifier



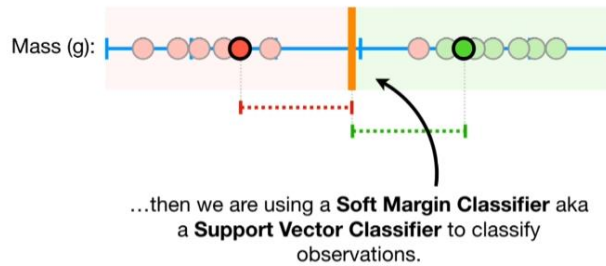
解决这个 Maximal Margin High Variance 的问题，可以忽略 Outlier Data，来做出预测。这个方法也是 Bias Variance Tradeoff 的，就是舍弃 Outlier Data，如上图，选出比较集中的 Data 来做 Threshold 计算，而这个算出来的 Margin 叫做 Soft Margin。



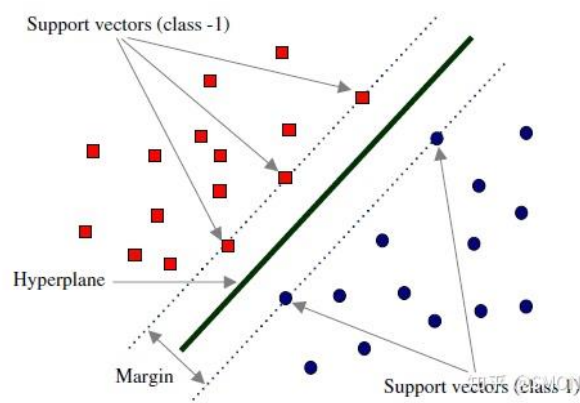
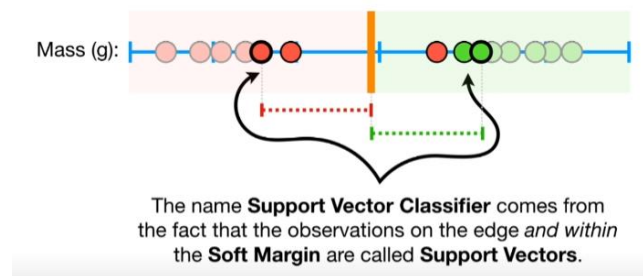
如上图，就不是一个最好的 Soft Margin。选择适合的 Soft Margin 可以使用 Cross Validation 的方法，使用同样的 Data，设置不同的 Threshold 来选出错误最低的 Threshold。



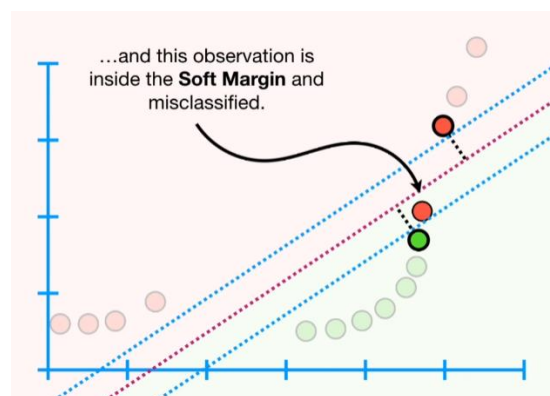
上图列子，如果上图的是 Cross Validation 后选出来的最好的 Soft Margin，那么在 Soft Margin 里会允许 1 个错误分类和 2 个正确分类。



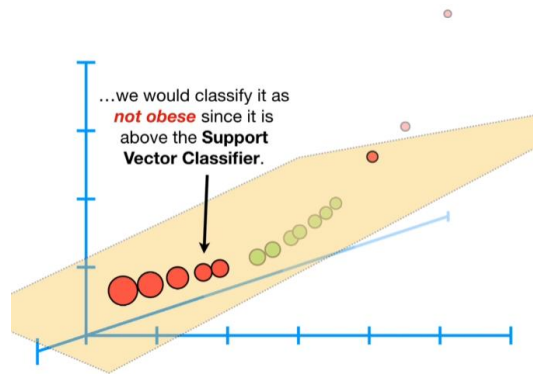
当使用 Soft Margin 来选出 Threshold 值的时候，也就是使用了 Soft Margin Classifier aka Support Vector Classifier。



Support Vector 的意思是一个 Class 里面边边角角的数据 Points 又刚好在 Soft Margin 里面，被称之为 Support Vector。



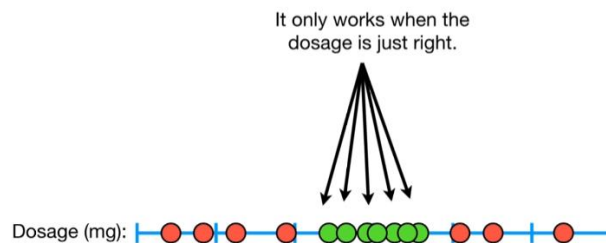
如上图，当有 2 个 Parameters 的时候，Soft Margin 是直线的 (1-Dimensional Subspace)。可以从上图看到在 Soft Margin 里，有一个错误分类的数据。通过 Cross Validation，发现放弃上图那个点可以让整个分类的工作表现更好。



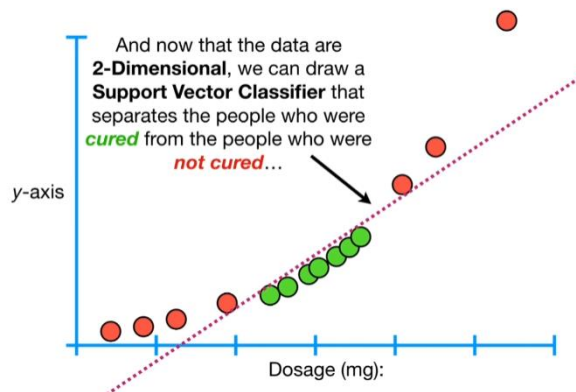
当 Parameters 是三维的时候，划分的就是一个 Plane (2-Dimensional Subspace)。Data Point 在 Plane 上面就是一个 Class，在 Plane 下面就是另一个 Class。

当 Parameters 是 4 维以上，这个 Support Vector Classifier 就会有一个 Hyperplane (Flat Affine Subspace)。

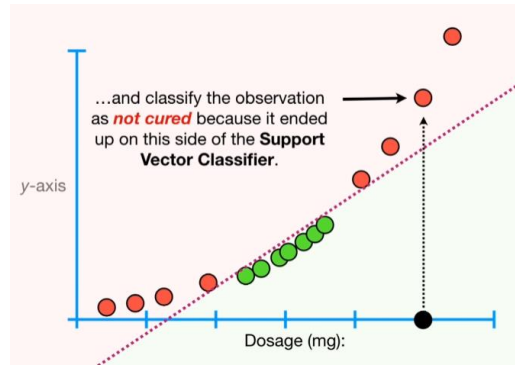
Support Vector Machine



如上图，当数据只有中间一部分是同一个 Class，而左右两边是另一个 Class，只是后 Maximal Margin Classifier 和 Support Vector Classifier 就不能很好的解决这个问题。这时候就需要用到 Support Vector Machine 来解决。

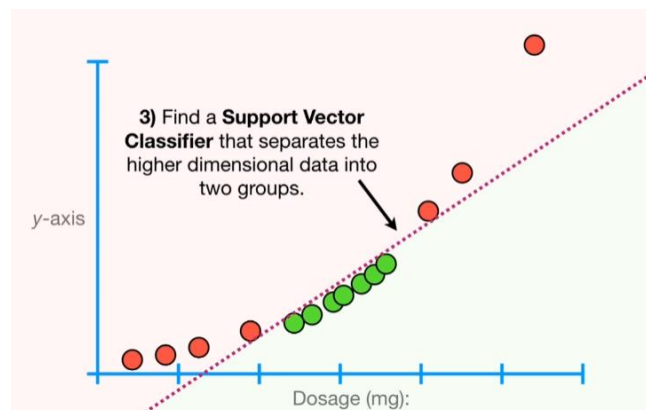
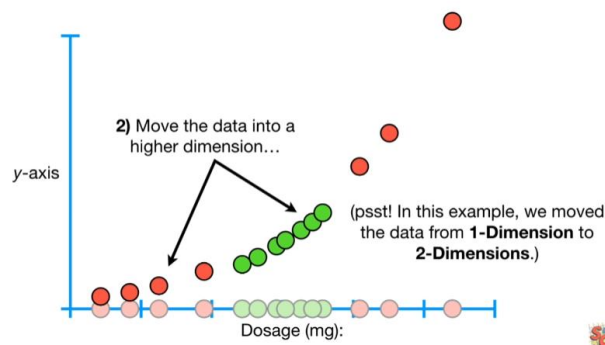
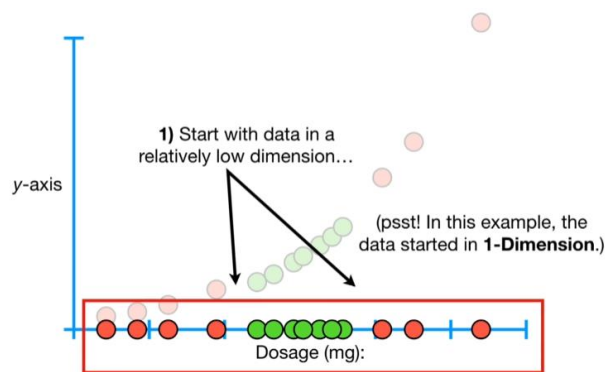


Support Vector Machine 的方法如上图，设置一个新的 y-axis，而这个 y-axis 是 Input Parameter 的平方， $Dosage^2$ ，当计算完之后就会得到一个二维的数据，只是后就能用直线来划分两个不同的 Class 如上图。



当 Support Vector Machine 要给出预测时候，新的 Input 值给如之后，取平方然后通过设定好的直线来分类，如上图。

Support Vector Machine Idea



Kernel Function

上面的例子，Y-axis 使用了 $Dosage^2$ ，但是为什么是取平方而不是其他的 Equations。在 SVM 里面，有一个东西叫做 Kernel Functions，是用来系统地找出 Support Vector Classifiers 在 Higher Dimensions。当计算 High-Dimensional Relationship without Actually Transforming the Data to the Higher-Dimension，也叫做 The Kernel Trick。

Polynomial Kernel

Polynomial Kernel 是用来计算两个 Observations (Data Points) 之间的关系。

$$\text{Polynomial Kernel} = (a * b + r)^d$$

$a \& b \rightarrow 2 \text{ Different Observations in Dataset}$

$r \rightarrow \text{Coefficient of Polynomial}$

$d \rightarrow \text{Degree of Polynomial}$

$r \& d$ 的值需要通过 Cross Validation 来认证

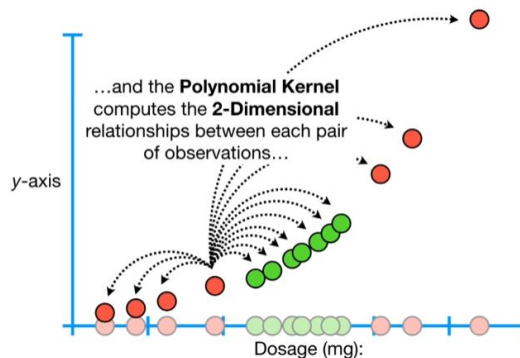
在这里 a 和 b 也代表着需要计算 High-Dimensional Relationship 值的 Observations。

基础概念

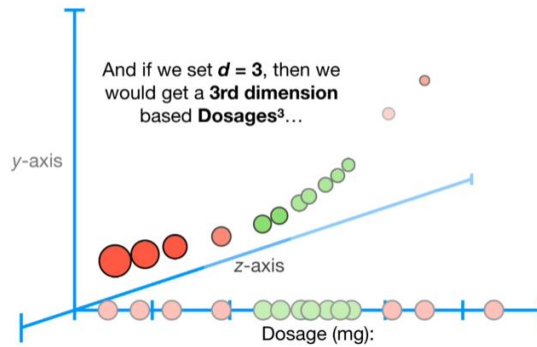
When $d = 1$, the **Polynomial Kernel** computes the relationships between each pair of observations in **1-Dimension**...



当 $d = 1$ 的时候，Kernel 会计算每一个 Data Point 与其他所有 Data Points 之间的距离，这是为了找出 Support Vector Classifier。



当 $d = 2$ 时，数据的 y-axis 会是 x-axis 的平方，然后也是会计算每一个 Data Point 与其他 Data Points 的距离，来找出 Support Vector Classifier。



当 $d = 3$ 时，第三个 Dimension 将会是 $Parameter^3$ 。然后再找出 Support Vector Classifier。 d 越大，Dimensions 越多。可以通过 Cross Validation 的方法来找出最适合的 d 值。

公式解释

当 $r = \frac{1}{2}$ 和 $d = 2$ 的时候

$$\text{Polynomial Kernel} = \left(a * b + \frac{1}{2}\right)^2$$

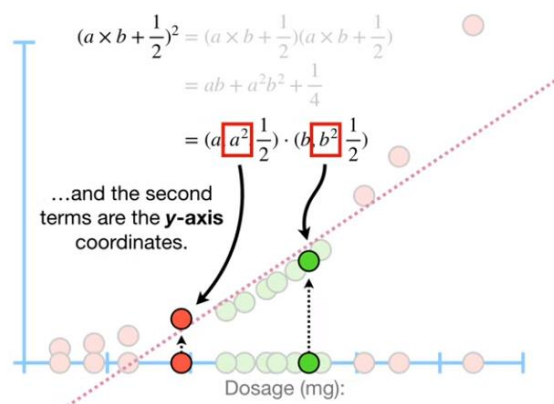
$$\left(a * b + \frac{1}{2}\right)^2 = \left(a * b + \frac{1}{2}\right) \left(a * b + \frac{1}{2}\right)$$

$$\left(a * b + \frac{1}{2}\right) \left(a * b + \frac{1}{2}\right) = a^2 b^2 + \frac{1}{2} ab + \frac{1}{2} ab + \frac{1}{4}$$

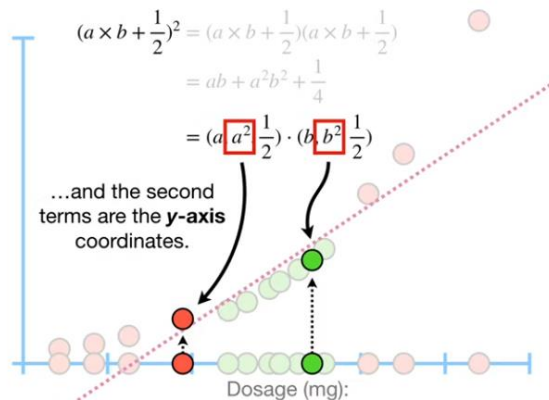
$$a^2 b^2 + \frac{1}{2} ab + \frac{1}{2} ab + \frac{1}{4} = ab + a^2 b^2 + \frac{1}{4}$$

$$ab + a^2 b^2 + \frac{1}{4} = \left(a, a^2, \frac{1}{2}\right) \cdot \left(b, b^2, \frac{1}{2}\right)$$

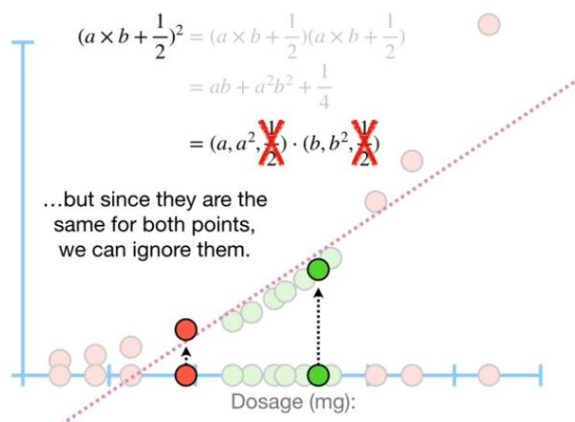
推算后，就是第一个 Observation 与第二个 Observation 之间取 Dot Product。Dot Product 能够让我们知道 High-Dimensional Coordinates for the data。



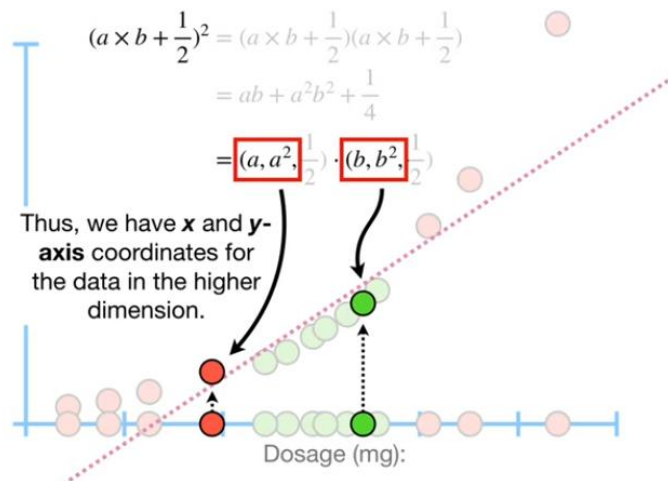
Dot Product 里的第一个值其实就是 Input Parameter 的值。



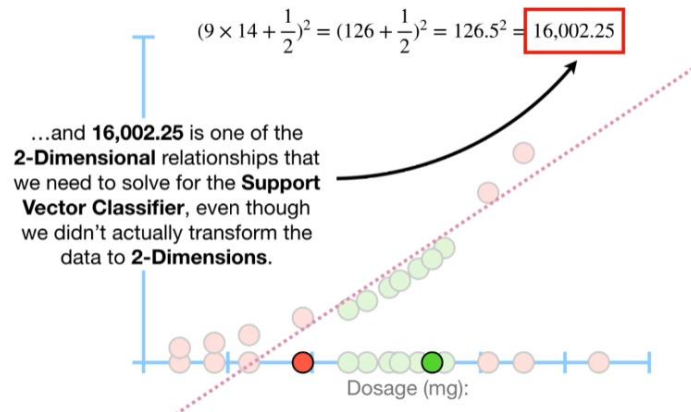
Dot Product 里的第二个值就是 Input Parameter 的平方。



在这里，Dot Product 的第三个值是 z-axis 的值，在这里两个值都一样，在这里可以不看。

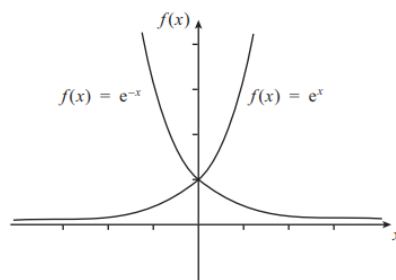


如上图，其实就得到了不同维度的值。所以只需要 Input 两个不同的 Data Point 的值，就能计算出 High-Dimensional Relationship 值。而 High-Dimensional Relationship 值就能用来找出 Support Vector Classifier。



上图给出的例子，两个 Data Point 之间的高维关系，计算出后得到了 16002.25，而这个值能够用来找出 Support Vector Classifier。使用这个方法的好处是不需要实质上的把 Data 转换成高纬度。

Radial Kernel aka Radial Basis Function (RBF)



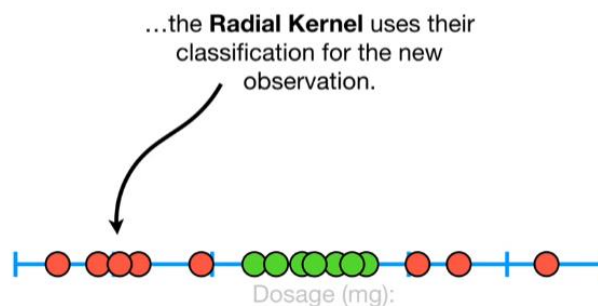
$$\text{Radial Kernel} = e^{-\gamma(a-b)^2}$$

a & $b \rightarrow$ Two Different Observations

$\gamma \rightarrow$ Scalar of the Squared Distance thur Cross Validation

基础概念

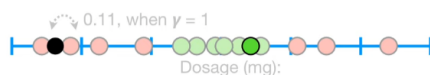
Radial Kernel 就好像 Weighted Nearest Neighbor，Radial Kernel 的概念就是计算出每一个 Data 对目前 Input 的影响力，越靠近目前 Input 的 Data 影响力越大，计算完所有 Data 之后，选出有着最大影响力的 Data 的 Class 当作目前 Input 的 Class。



...and when the points are relatively far from each other, we get **A Number Very Close to Zero.**

$$e^{-(2.5-16)^2} = e^{-(-13.5)^2} = e^{-182.25}$$

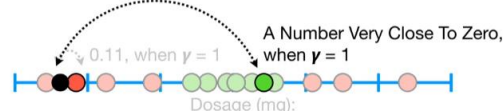
A Number = Very Close to Zero.



从上图可以看到，当目标靠近的时候， $Output = 0.11$ ，而当目标远的时候，算出来的值接近 0，也就是 Less Influence。

...and **A Number Very Close to Zero** is the high-dimensional relationship between these two observations that are relatively far from each other.

$$e^{-\gamma(a-b)^2} = \text{high-dimensional relationship}$$



当使用 Radial Kernel 来计算影响力的时候，其实就是计算 High Dimensional Relationship between these Two Observations。

$$\text{When } r = 0 \dots (a \times b + r)^d = (a \times b)^d = a^d b^d = (a^d) \cdot (b^d)$$

$$\text{When } d = 1 \text{ we get } \dots a^1 b^1 = (a) \cdot (b)$$

$$\text{When } d = 2 \text{ we get } \dots a^2 b^2 = (a^2) \cdot (b^2)$$

$$\text{When } d = 3 \text{ we get } \dots a^3 b^3 = (a^3) \cdot (b^3)$$

So, setting $r = 0$ seems silly because no matter what values we use for d , the **Dot Products** leave the data in the original dimension...



在使用 Polynomial Kernel 的时候，当 $r = 0$ 的时候，算出来的值就只是 1-Dimensional 的直线值。如上图，不管 d 的值再怎么变，算出来的值只是把原来的 Observation 值 Scale 了。但是在 $r = 0$ 的情况下，不停的加不同 d 的公式进来，就能计算出不同维度的值。

$$\text{When } r = 0 \dots (a \times b + r)^d = (a \times b)^d = a^d b^d = (a^d) \cdot (b^d)$$

$$a^1 b^1 + a^2 b^2 = (a, a^2) \cdot (b, b^2)$$

This gives us a **Dot Product** with coordinates for **2-Dimensions**.




当 $d = 1$ 加上了 $d = 2$ 的 Polynomial Kernel，就能得到 2-Dimensional 的 Dot Product。

When $r = 0 \dots (a \times b + r)^d = (a \times b)^d = a^d b^d = (a^d) \cdot (b^d)$

$$a^1 b^1 + a^2 b^2 + a^3 b^3 + \dots + a^\infty b^\infty = (a, a^2, a^3, \dots, a^\infty) \cdot (b, b^2, b^3, \dots, b^\infty)$$

That would give us a **Dot Product** with coordinates for an *infinite number of dimensions!!!!*



如果持续不停的加，最终就会得到 Infinite Number of Dimension，这个举动正是 Radial Kernel 在做的。

算法推导

$$\text{Radial Kernel} = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-ab)}$$

$$e^{-\gamma(a^2+b^2-ab)} = e^{-\gamma(a^2+b^2)} e^{\gamma 2ab}$$

Taylor Series Expansion

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots + \frac{f^\infty(a)}{\infty!}(x-a)^\infty$$

Taylor Series 想法是 把一个 Function Split 成 Infinite Sum，如上面的这个 Equation，计算出的值是接近与 Function 但是不等于 Function。Taylor Series says a can be any value as long as $f(a)$ exists。

$$\frac{d}{dx} e^x = e^x$$

$$e^x = e^a + \frac{e^a}{1!}(x-a) + \frac{e^a}{2!}(x-a)^2 + \dots + \frac{e^a}{\infty!}(x-a)^\infty$$

Since $e^0 = 1$ exists, let set $a = 0$

$$e^x = e^0 + \frac{e^0}{1!}(x-0) + \frac{e^0}{2!}(x-0)^2 + \dots + \frac{e^0}{\infty!}(x-0)^\infty$$

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \dots + \frac{1}{\infty!}x^\infty$$

Solve for $e^{\gamma 2ab}$

$$e^{ab} = 1 + \frac{1}{1!}(ab) + \frac{1}{2!}(ab)^2 + \dots + \frac{e^0}{\infty!}(ab)^\infty$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \dots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

Adding up bunch of Polynomial Kernels with $r = 0$ and d from 0 to infinity

$$a^0 b^0 + a^1 b^1 + a^2 b^2 + \dots + a^\infty b^\infty = (1, a^1, a^2, \dots, a^\infty) \cdot (1, b^1, b^2, \dots, b^\infty)$$

...and a **Polynomial Kernel** with $r = 0$ and $d = 2$ is equal to $(ab)^2$...

$$a^0 b^0 + a^1 b^1 + \boxed{a^2 b^2} + \dots + a^\infty b^\infty = (1, a, a^2, \dots, a^\infty) \cdot (1, b^1, b^2, \dots, b^\infty)$$

$$e^{ab} = 1 + \frac{1}{1!}ab + \frac{1}{2!}\boxed{(ab)^2} + \frac{1}{3!}(ab)^3 + \dots + \frac{1}{\infty!}(ab)^\infty$$

从上图能发现，当 Polynomial Kernel $r = 0$ 时，也就是跟 Radial Kernel 在做一样的事情。

Back to Radial Kernel Equation

$$\text{Radial Kernel} = e^{-\gamma(a^2+b^2)} e^{\gamma 2ab}$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^2, \dots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^2, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

Let $\gamma = \frac{1}{2}$

$$\begin{aligned} \text{Radial Kernel} &= e^{-\frac{1}{2}(a^2+b^2)} e^{ab} \\ &= e^{-\frac{1}{2}(a^2+b^2)} \left[\left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^2, \dots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^2, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right) \right] \end{aligned}$$

Further Simplify

$$s = \sqrt{e^{-\frac{1}{2}(a^2+b^2)}}$$

$$e^{-\frac{1}{2}(a^2+b^2)} = \left(s, s\sqrt{\frac{1}{1!}}a, s\sqrt{\frac{1}{2!}}a^2, s\sqrt{\frac{1}{3!}}a^2, \dots, s\sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(s, s\sqrt{\frac{1}{1!}}b, s\sqrt{\frac{1}{2!}}b^2, s\sqrt{\frac{1}{3!}}b^2, \dots, s\sqrt{\frac{1}{\infty!}}b^\infty\right)$$

...and, at long last, we see that the **Radial Kernel** is equal to a **Dot Product** that has coordinates for an infinite number of dimensions.

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2)} \left[\left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \dots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right) \right]$$

$$e^{-\frac{1}{2}(a-b)^2} = \left(s, s\sqrt{\frac{1}{1!}}a, s\sqrt{\frac{1}{2!}}a^2, \dots, s\sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(s, s\sqrt{\frac{1}{1!}}b, s\sqrt{\frac{1}{2!}}b^2, \dots, s\sqrt{\frac{1}{\infty!}}b^\infty\right)$$

从上面的公式推导，和上图，可以看出当在计算 $e^{-\gamma(a^2+b^2)}$ ，其实就是在计算两个点之间的 Infinite-Dimensions 的关系。