Feature Engineering Example

Paul Pearson

December 19, 2015

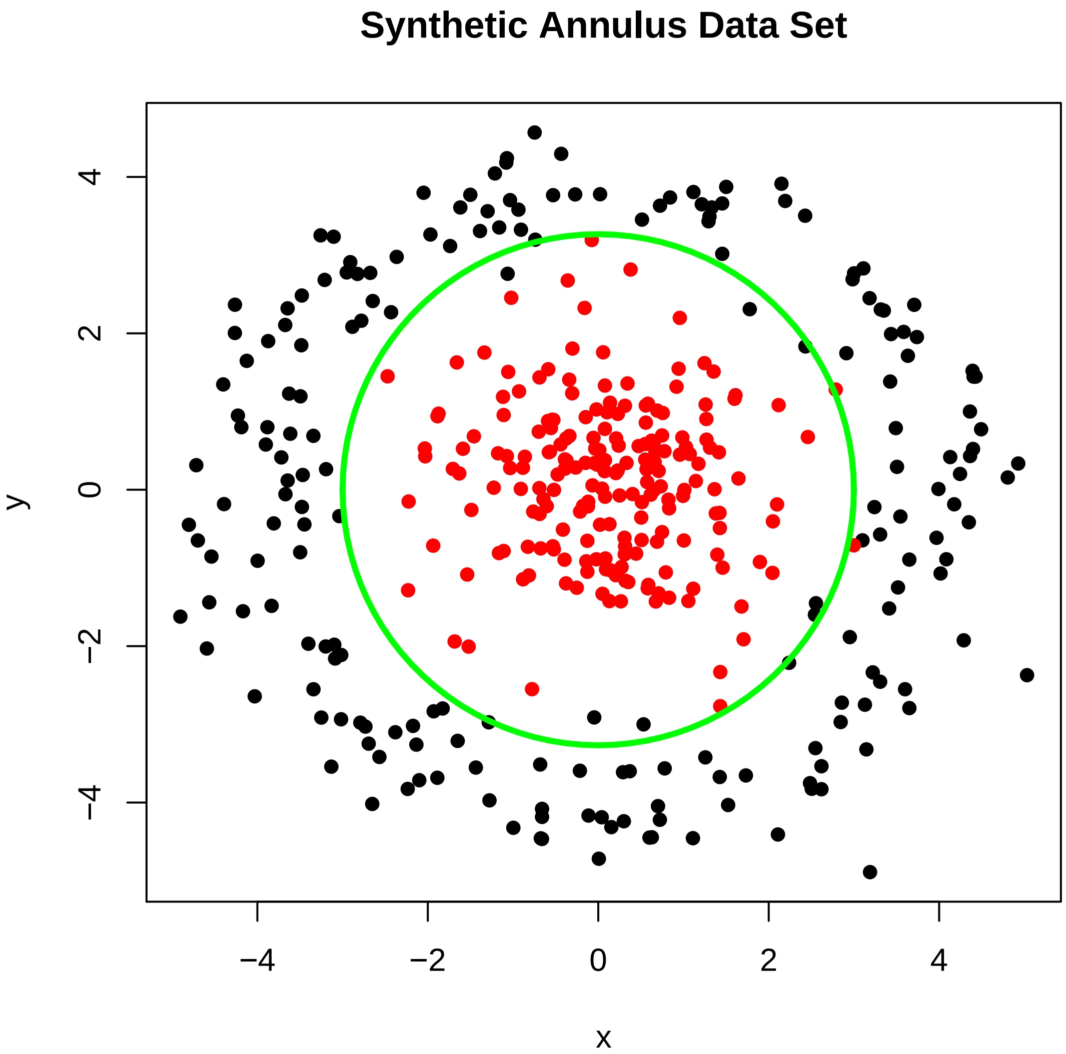
Construct an artificial data set

Let's create an artificial data set consisting of randomly generated (black) points in the shape of an annulus centered at the origin, together with randomly generated (red) points that mostly lie within the annulus. These points are **not** "linearly separable" – there is no single line that will separate the red points from the black points.

```r
# 200 random points on an "annulus"
n <- 200 # number of points in annulus
t <- runif(n, min = 0, max = 2*pi) # random uniform distribution
r <- rnorm(n, mean = 4, sd = 0.5) # random gaussian (normal) distribution
x1 <- r*cos(t)
y1 <- r*sin(t)
z1 <- rep(1,n) # class label = 1

# 200 random points inside the "annulus"
x2 <- rnorm(n, mean = 0, sd = 1)
y2 <- rnorm(n, mean = 0, sd = 1)
z2 <- rep(2,n) # class label = 2

library(plotrix)
plot(c(x1,x2), c(y1,y2),
     type="p", pch=16, # plot points with filled circle marker
     col=c(z1,z2), # color points according to class labels
     main = "Synthetic Annulus Data Set",
     xlab = "x",
     ylab = "y")
draw.circle(x=0, y=0, radius=3, border="green", lwd=3)
```



Notice that the green circle of radius 3, which is not part of the data, does a pretty good job of separating the red and black points.

Feature engineering

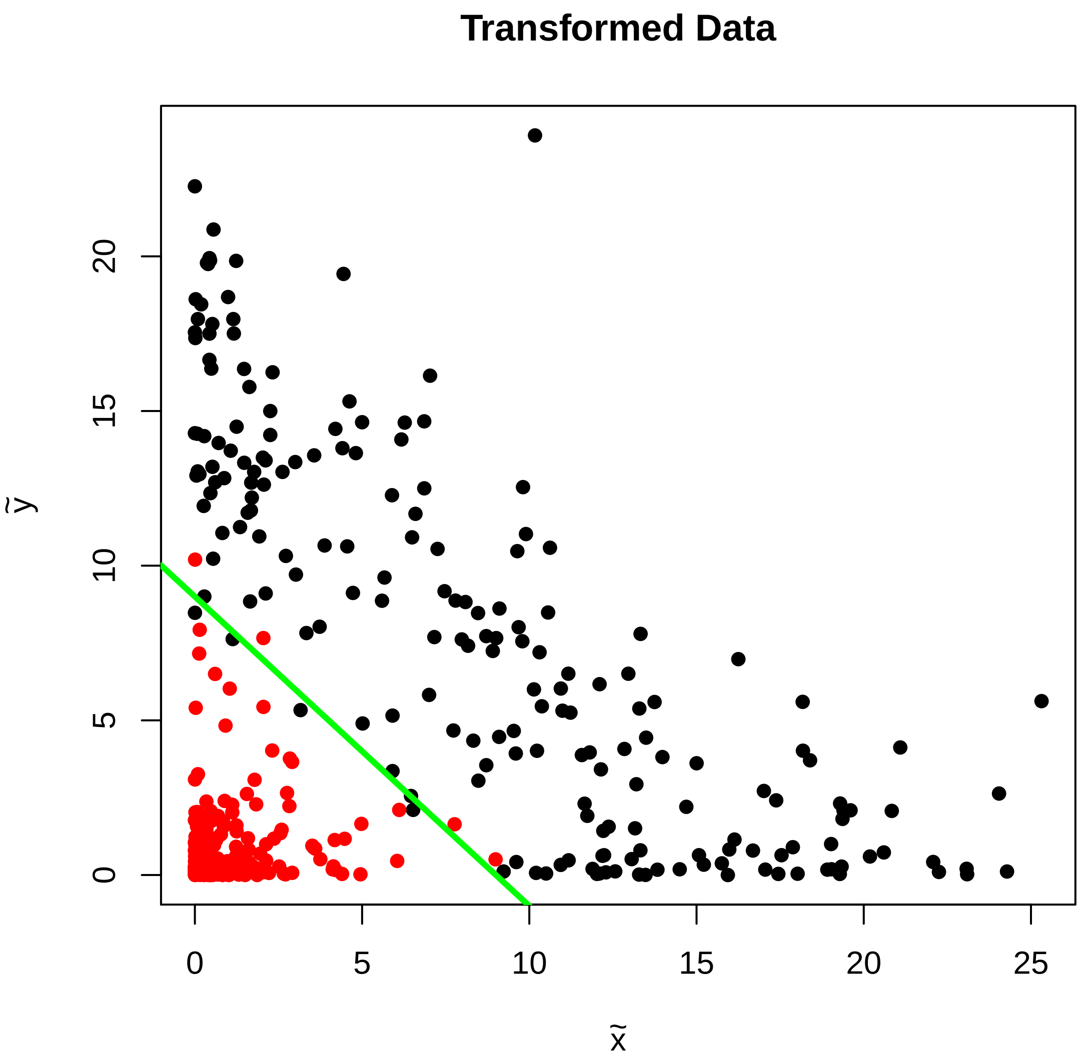
Can we transform the given data points (x, y) into points $(\tilde{x}, \tilde{y})$ that are linearly separable? In other words, can we find functions $\tilde{x} = f(x, y)$ and $\tilde{y} = g(x, y)$ so that the points $(\tilde{x}_i, \tilde{y}_i) = (f(x_i, y_i), g(x_i, y_i))$ are linearly separable? If we use the functions $\tilde{x} = f(x, y) = x^2$ and $\tilde{y} = g(x, y) = y^2$, something very interesting happens under the transformation $(x, y) \mapsto (x^2, y^2)$.

- Thought question: Consider the set of all points on a circle of radius r (i.e., satisfying $x^2 + y^2 = r^2$). Choose a radius r and several points (x_i, y_i) on a circle of radius r. What are the images of these points under the transformation $(x, y) \mapsto (x^2, y^2)$ that squares each coordinate.
- Thought question: Consider the set of all points on a line through the origin (such as the x-axis, the y-axis, and the line $y = mx$). What is the image of these points under the transformation $(x, y) \mapsto (x^2, y^2)$

that squares each coordinate.

- Thought question: What will happen to the synthetic annulus data under the transformation $(x, y) \mapsto (x^2, y^2)$? The result is shown in the graph below.

```
library(latex2exp)
plot(c(x1^2,x2^2), c(y1^2,y2^2),
     type="p", pch=16, # plot points with filled circle marker
     col=c(z1,z2), # color points according to class labels
     main = "Transformed Data",
     xlab = TeX('$\\tilde{x}$'),
     ylab = TeX('$\\tilde{y}$'))
abline(a=9, b=-1, col="green", lwd="3")
```



Notice that under the transformation $(x, y) \mapsto (x^2, y^2) = (\tilde{x}, \tilde{y})$, the points on a circle of radius r in xy-space get transformed into points on the line $\tilde{y} = r\tilde{x}$ in the first quadrant in $\tilde{x}\tilde{y}$-space. So, now in $\tilde{x}\tilde{y}$-space we can separate the red and black points using a single line! The transformation has turned a problem in xy-space that is not linearly separable into a problem in $\tilde{x}\tilde{y}$-space that is linearly separable by the green line! So, when $\tilde{x} + \tilde{y} \leq 9$, the point with coordinates $(\tilde{x}, \tilde{y})$ is very likely to be red.

So, as long as we can find the center of the annulus, we can linearly separate things inside the annulus from things outside the annulus using the transformation $(x, y) \mapsto (x^2, y^2)$.