What's the point?

$$\theta_j^{(i+1)} \leftarrow \theta_j^{(i)} + \alpha \frac{\delta}{\delta \theta_j} J(\theta^{(i)})$$

Model parameters can be anything

Cost function can be anything*

*needs to be differentiable

Other cost functions

vector of weights

vector of instance data

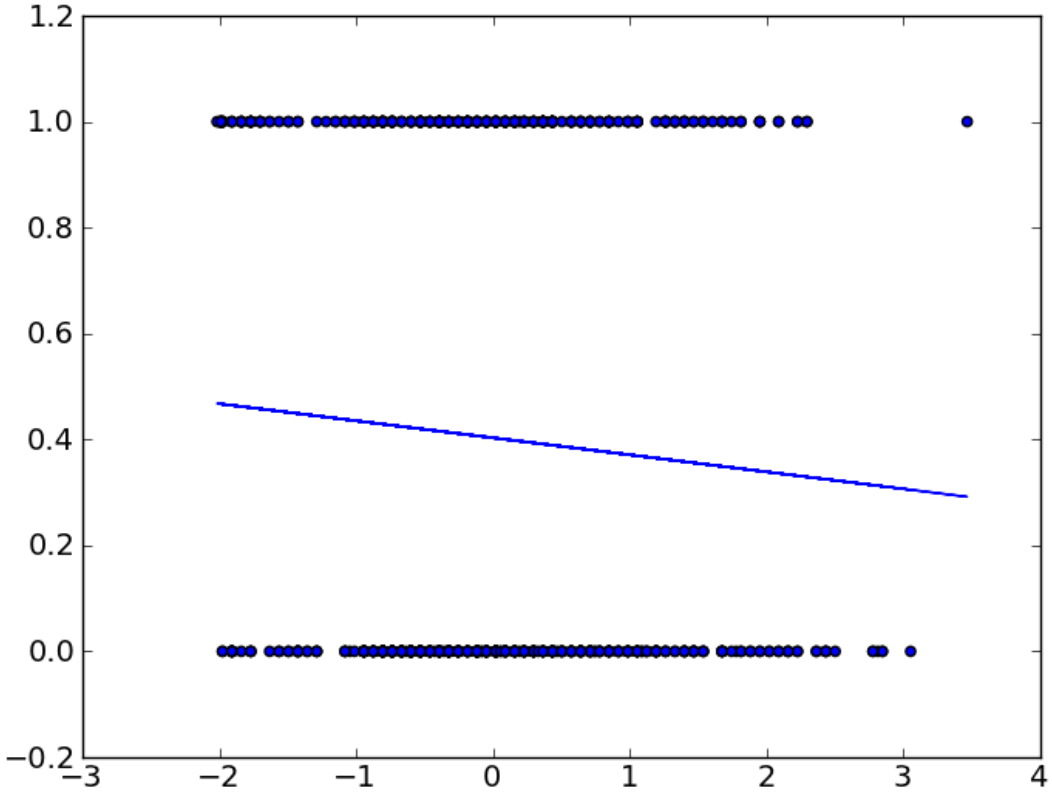
Logistic Regression

$$J(\theta) = \frac{1}{n}\sum_{i=0}^{n} \log_2\left(1 + \exp(-y_i(\theta \cdot x_i))\right) + \frac{\lambda}{2}||\theta||^2$$

Support Vector Machines

$$J(\theta) = \frac{1}{n}\sum_{i=0}^{n} \max(1 - y_i(\theta \cdot x_i), 0) + \frac{\lambda}{2}||\theta||^2$$

Quick Intuition for Logistic Regression (1)



Gradient descent regression line for titanic data.

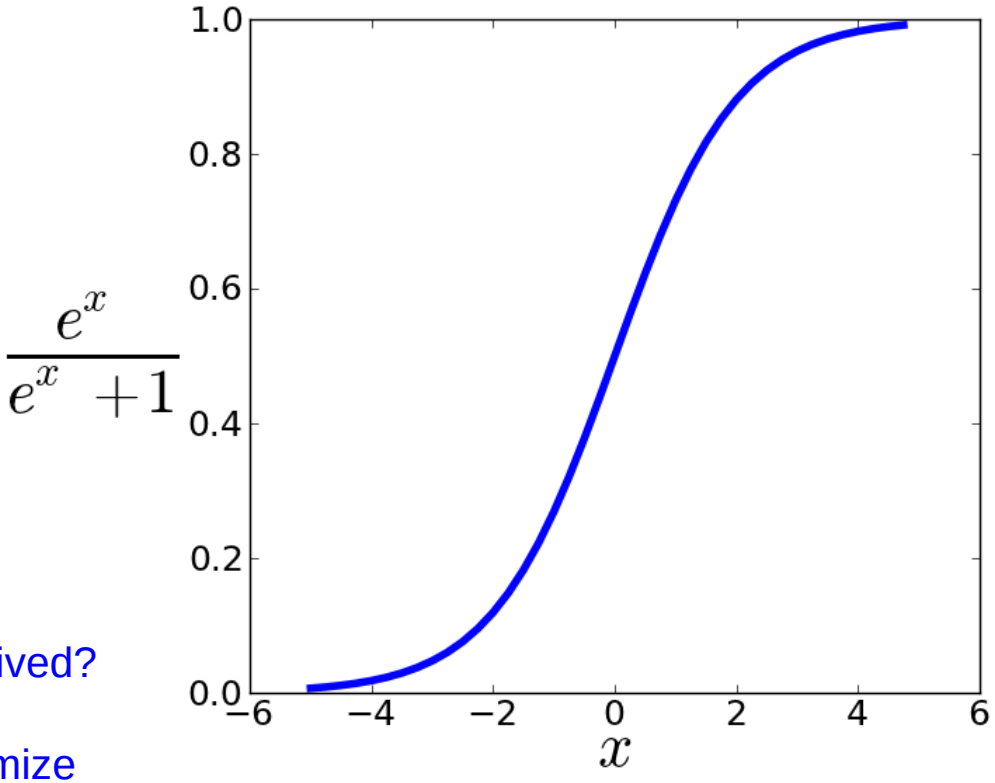
Predicting survival (y-axis) from (normalized) age (x-axis).

What does this line mean?

Nothing really

Quick Intuition for Logistic Regression (2)

$$f(x) = \frac{e^x}{e^x + 1}$$



Maps **any** number to the range (0,1)

Interpret the result as a probability

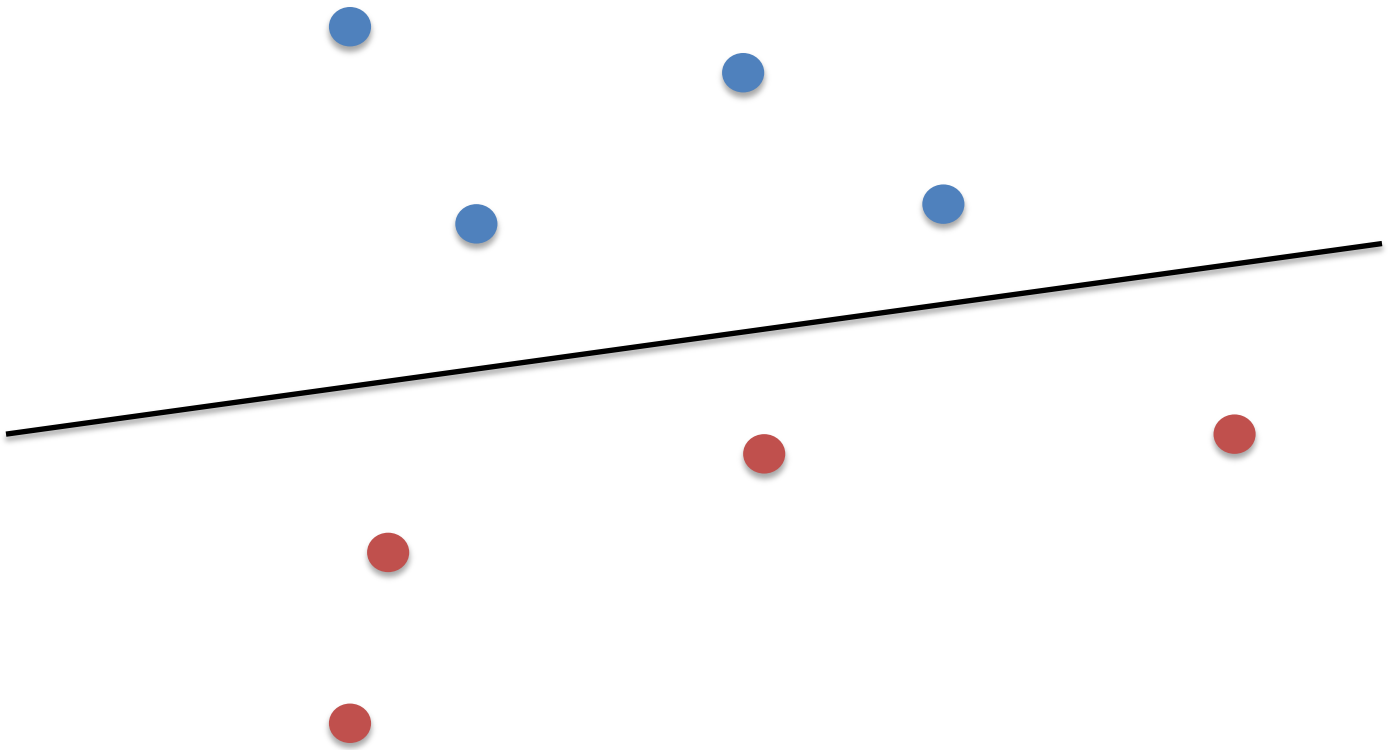
Interpret categorical classes numerically

What is the *probability* a passenger survived?

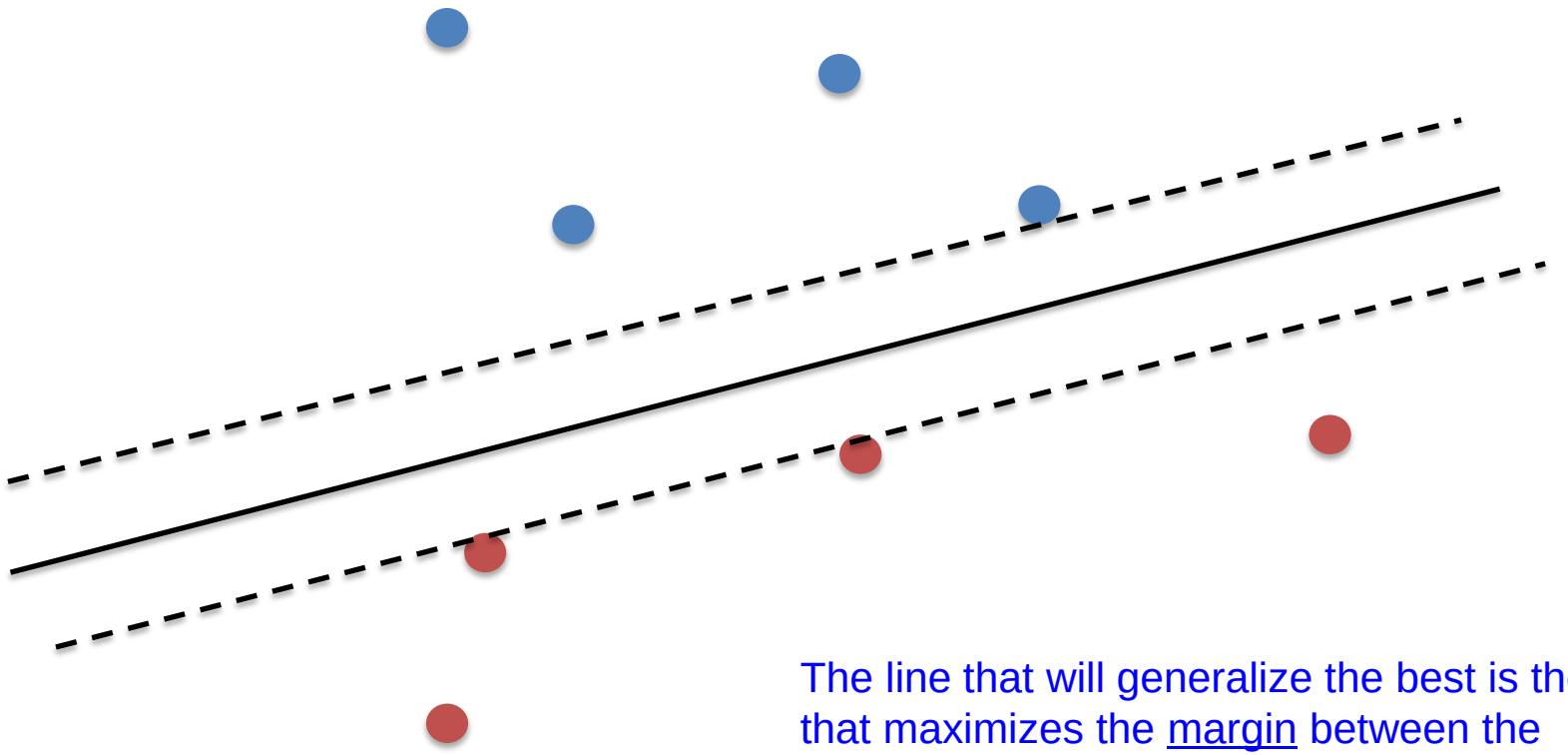
The cost function is constructed to maximize the probability of correct classification.

Easy to work with

Quick Intuition for Support Vector Machines (1)

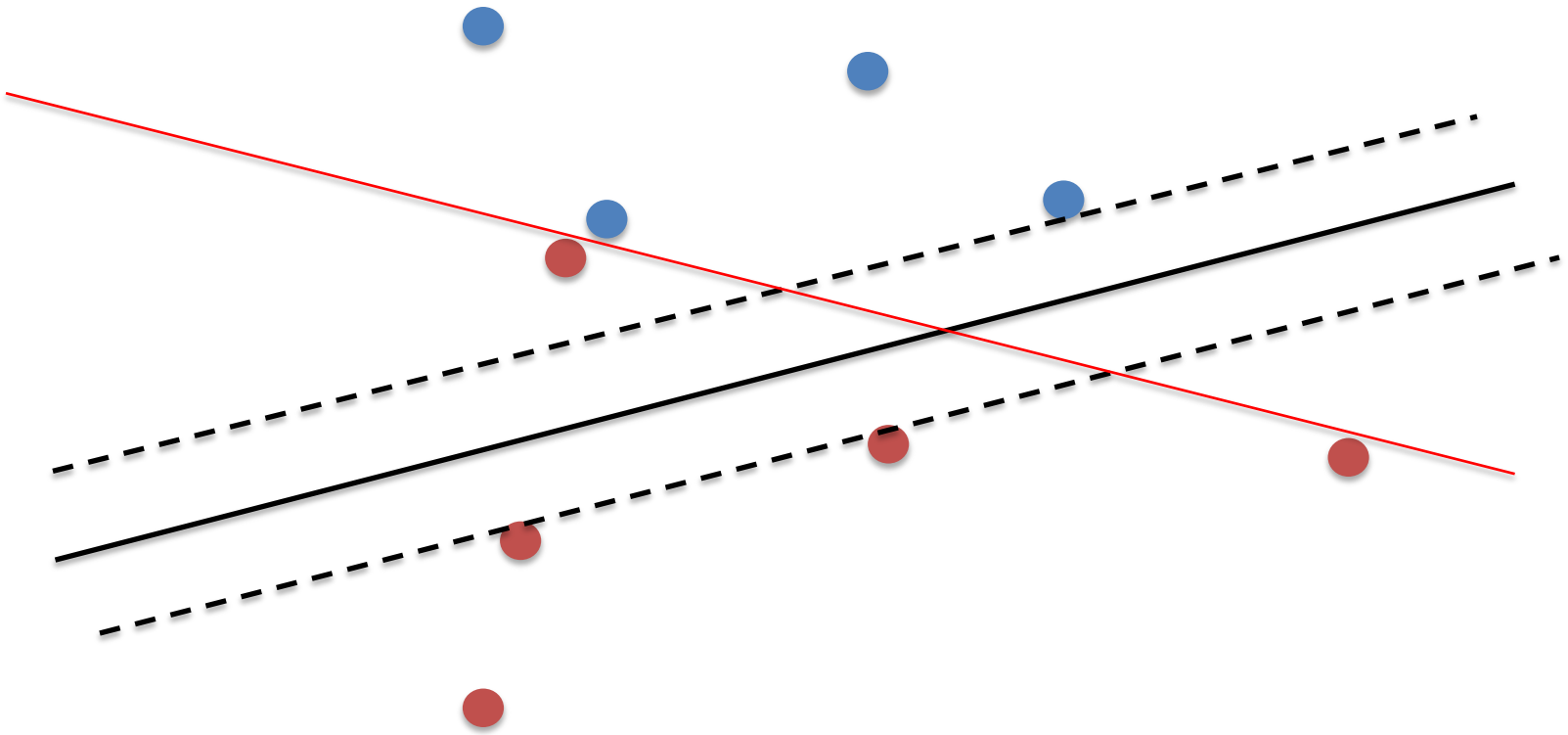


Quick Intuition for Support Vector Machines (2)



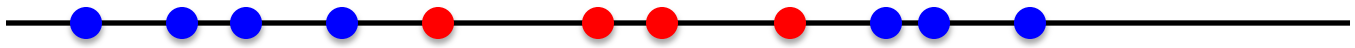
The line that will generalize the best is the one that maximizes the margin between the classes.

Quick Intuition for Support Vector Machines (3)

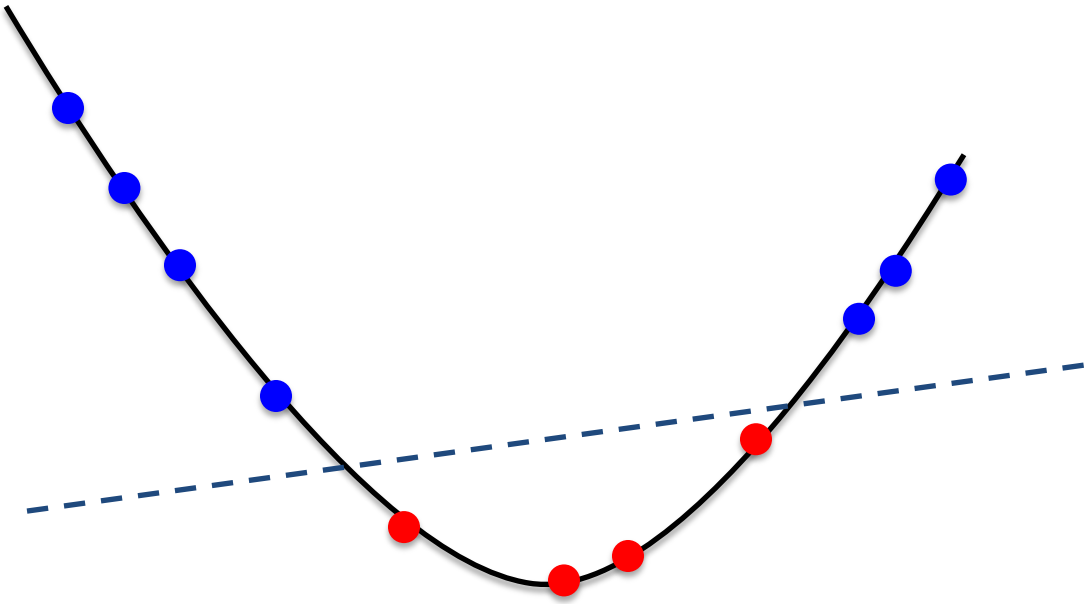


Quick Intuition for Support Vector Machines (4)

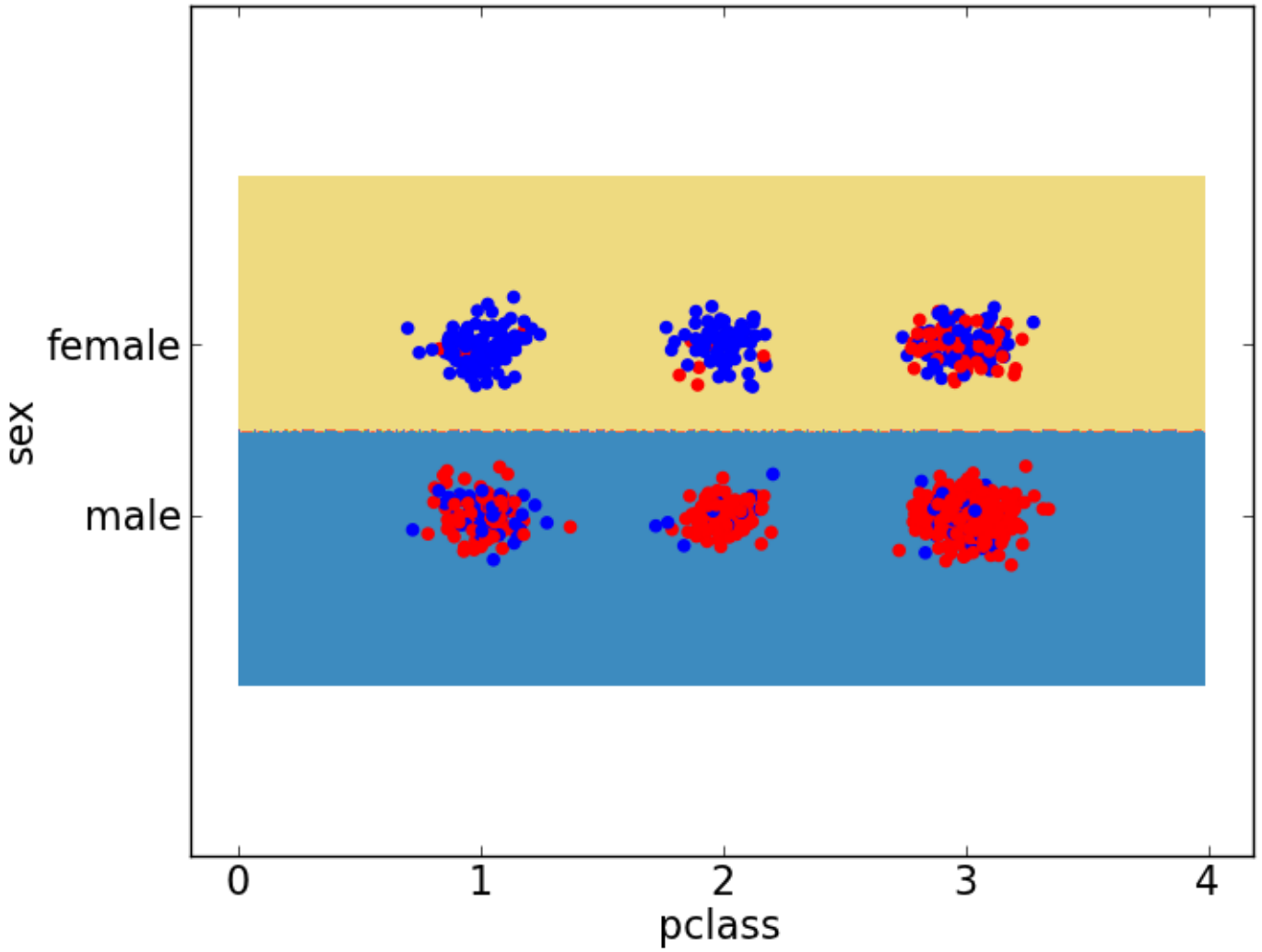
Not linearly separable in 1D:



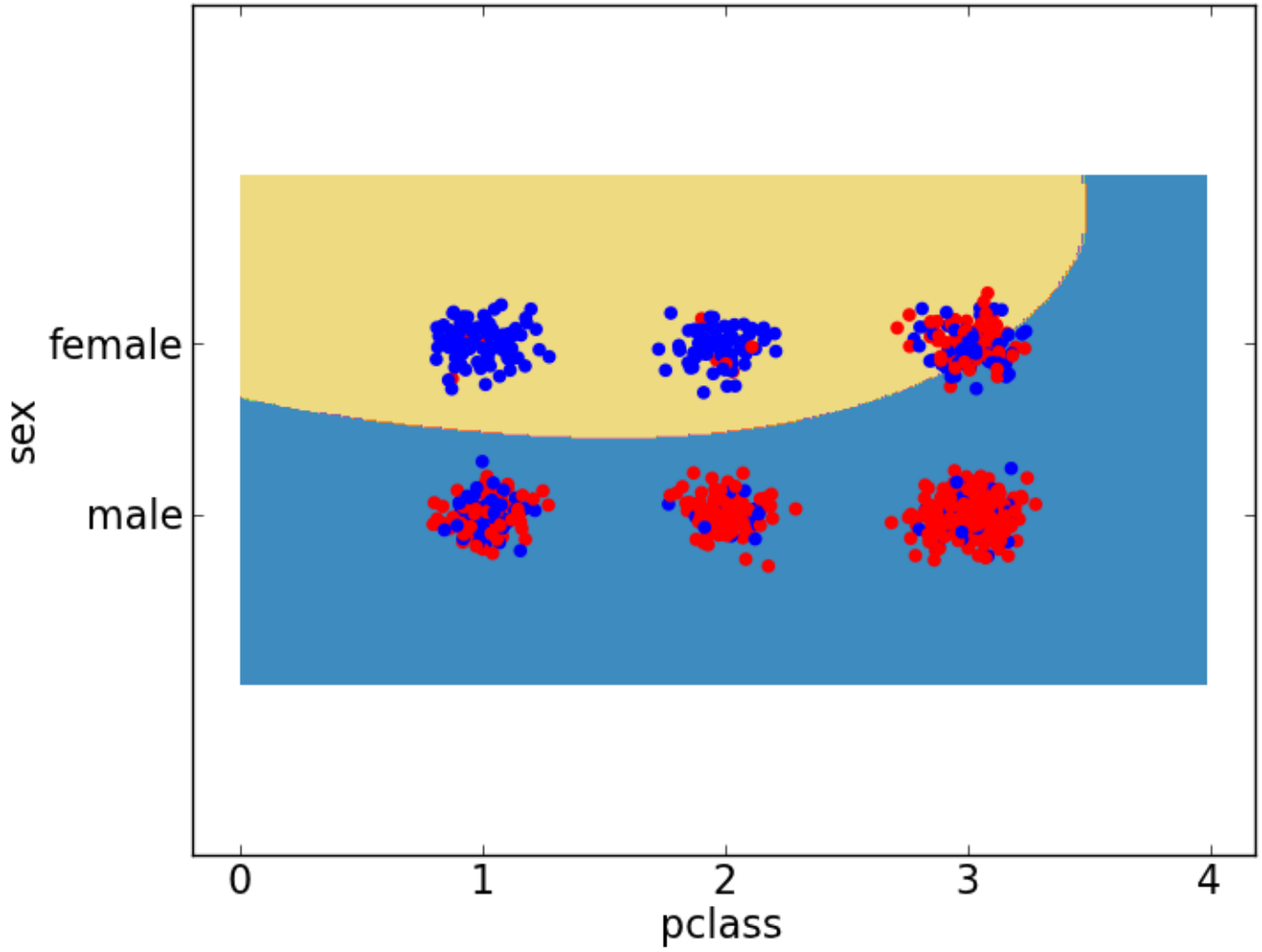
Map the data into a higher dimensional space by applying a kernel



Support Vector Machine Model, Titanic Data, Linear Kernel

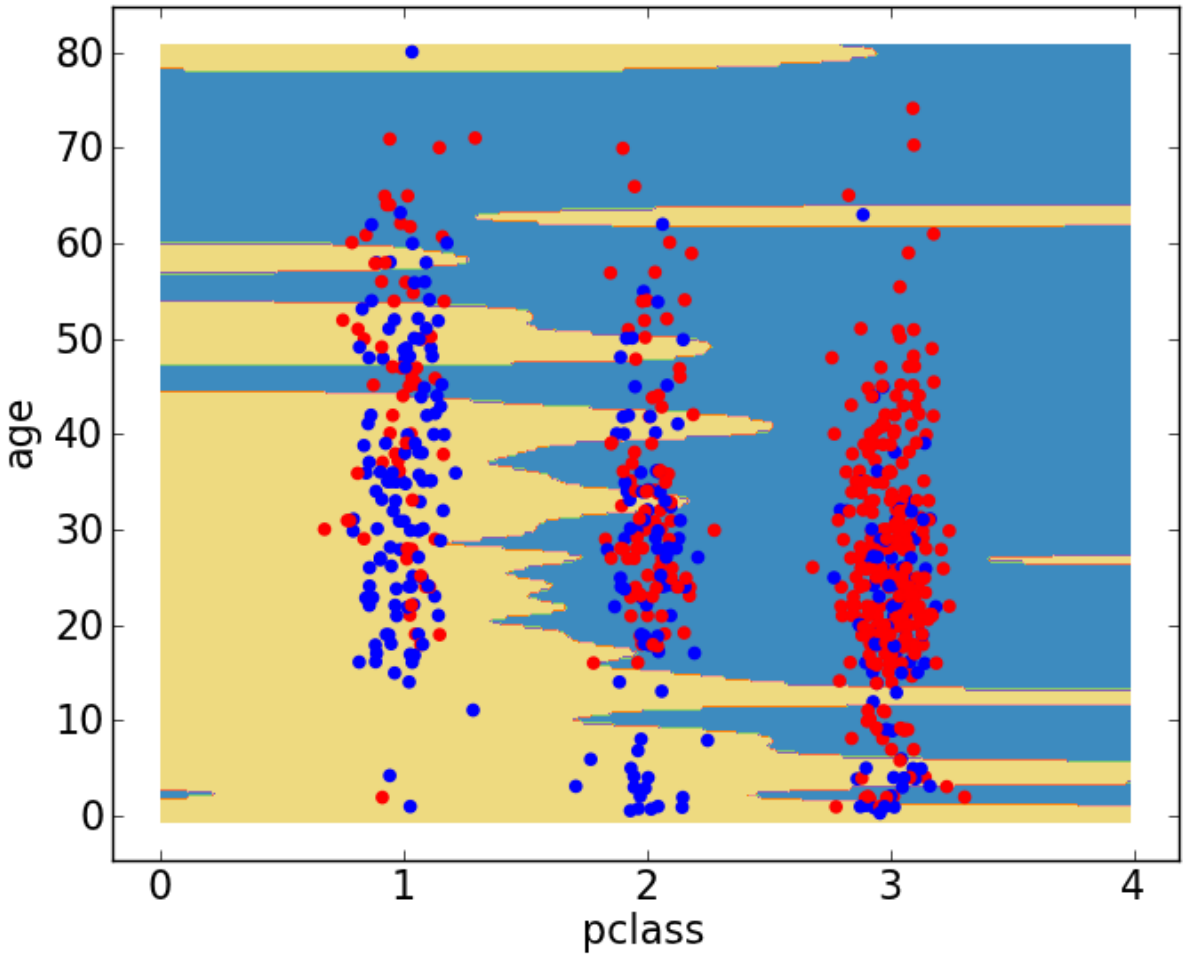


Support Vector Machine Model, Titanic Data, Radial Basis Function Kernel



Support Vector Machine Model, Titanic Data, Radial Basis Function Kernel

overfitting?



Support Vector Machine Model, Titanic Data, Radial Basis Function Kernel

a lower gamma, a parameter that controls that balances model complexity against accuracy