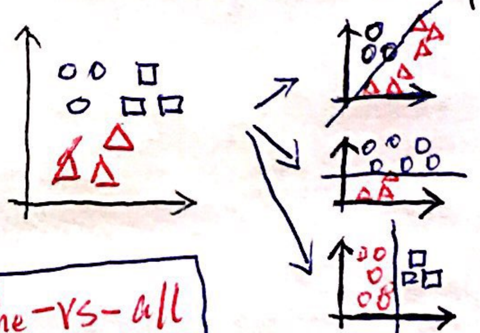
Multiclass Classification.



$\boxed{\text{One-vs-all}}$

if $y \in \{0, 1, 2, k\}$

Then train a logistic regression classifier $h_\theta(x)$ for each class to predict the probability that $y=i$.

$$h_\theta^{(0)}(x) = P(y=0 \mid x;\theta)$$

$$h_\theta^{(1)}(x) = P(y=1 \mid x;\theta)$$

$$\vdots$$

$$h_\theta^{(k)}(x) = P(y=k \mid x;\theta)$$

Finally, to make a prediction on a new x, pick the class that maximizes $h_\theta(x)$

$$\text{prediction} = \max_i \left(h_\theta^{(i)}(x)\right)$$

Addressing Overfitting

1. Reduce number of features.
   — select features to keep.
   — Model selection.
2. Increase number of traing data.
3. Regularization
   — keep all features, but reduce magnitude of θ_j.

# of features

$$Var \sim \frac{n}{m}$$

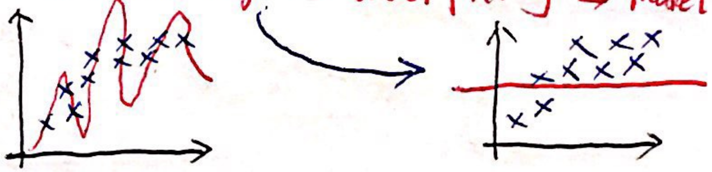
# of training data

Regularization.

$$\min_\theta \; J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

λ is too small → overfitting → no penalty to overfitting

λ is too large → underfitting → model is too simple.



only $\theta_0 \neq 0$

all $\theta_j \to 0$

Regularized Linear Regression.

Gradient descent:

same as original LR

Repeat {

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]$$

}

$$\theta_j := \underbrace{\left(1 - \alpha\frac{\lambda}{m}\right)}_{<1}\theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]x_j^{(i)}$$

compensation for regularization term

Normal Equation

$$X = \begin{bmatrix}(x^{(1)})^T \\ \vdots \\ (x^{(m)})^T\end{bmatrix} \quad y = \begin{bmatrix}y^{(1)} \\ \vdots \\ y^{(m)}\end{bmatrix}$$

$\mathbb{R}^{m\times(n+1)}$

$$\theta = \left(X^TX + \lambda\begin{bmatrix}0 & \\ & 1 & \\ & & \ddots \\ & & & 1\end{bmatrix}\right)^{-1} X^Ty$$

$\mathbb{R}^{(n+1)\times(n+1)}$

if $\lambda > 0$, $\lambda\begin{bmatrix}0 & \\ & 1 & \\ & & \ddots \\ & & & 1\end{bmatrix}$ make it always invertible.

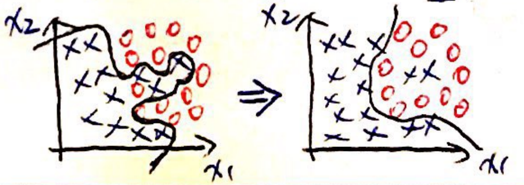
Regularized Logistic Regression.

$$h_\theta(x) = \frac{1}{1+e^{-\theta^Tx}}$$

Gradient descent:

Same as the Regularized Linear Regression. But $h_\theta(x)$ is different.

$$\min_\theta \; J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$



Oct 18, 2018