Universidade Federal de Santa Catarina
Centro Tecnológico
Departamento de Automação e Sistemas

Aprendizado supervisionado
Prof. Eric Antonelo

Representação de Modelo e Função de Custo

Representação de Modelo

A notação usada nesse curso será: $x^{(i)}$ para denotar variáveis de *entrada* (área do imóvel no exemplo), também chamadas de características de entrada (*features*), e $y_{(i)}$ para denotar variável de *saída* ou desejada (*target*), que estamos buscando estimar (preço do imóvel). Um par $(x^{(i)}, y^{(i)})$ é chamado um exemplo de treinamento, e o conjunto de dados que usaremos - um conjunto de m exemplos de treinamento, $\mathcal{D} = \{(x^{(i)}, y^{(i)}); i = 1, \cdots, m\}$ - é chamado de conjunto de treinamento. Note que o sobrescrito "(i)"na notação é simplesmente um índice no conjunto de treinamento e não tem nada a ver com exponenciação. Também usaremos X para indicar o espaço dos valores de entrada e Y para indicar o espaço dos valores de saída. Neste exemplo, X = Y = $\mathbb{R}$.

Para descrever o problema de aprendizado supervisionado de maneira um pouco mais formal, nosso objetivo é, dado um conjunto de treinamento, aprender uma função h: X $\rightarrow$ Y, de modo que h(x) seja um "bom"preditor para o valor correspondente de y. Por razões históricas, essa função h é chamada de *hipótese*. O processo é ilustrado a seguir:

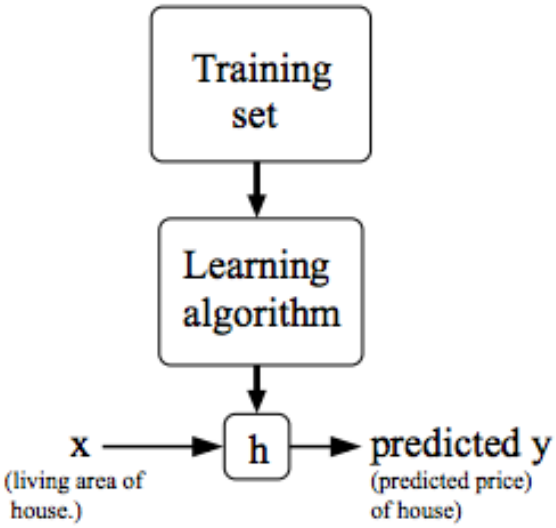

Figura 1: Treinamento de hipótese (modelo)

Quando a variável de saída que estamos buscando prever é contínua, como em nosso exemplo do imóvel, denominamos o problema de aprendizagem como um problema de **regressão**. Quando y pode assumir apenas um pequeno número de valores discretos (como se, dada a área de estar, desejássemos prever se uma habitação é uma casa ou um apartamento, por exemplo), chamamos isso de problema de **classificação**.

Função de Custo

Podemos medir a acurácia de nossa função *hipótese* usando uma **função de custo**. Ela mede a diferença média entre a saída da hipótese $h(x^{(i)})$, tendo como entrada o valor $x^{(i)}$ do i-ésimo exemplo, e a saída correspondente desejada $y^{(i)}$:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{1}$$

Essa função também é chamada de **erro quadrático médio**, ou *mean squared error*, e assume que a saída $y^{(i)}$ coletada no conjunto de treinamento $\mathcal{D}$ tem um valor constituído do valor verdadeiro $y_t^{(i)}$ (e inacessível) e um ruído Gaussiano $z \sim \mathcal{N}(0, \rho)$, i.e., $y^{(i)} = y_t^{(i)} + z$. A média é dividida pela metade $(\frac{1}{2})$ convenientemente para o cálculo do descenso do gradiente, pois o termo resultante da derivada da função potenciação (ao quadrado) cancelará com o termo $(\frac{1}{2})$.

A imagem seguinte resume o que a função de custo faz:

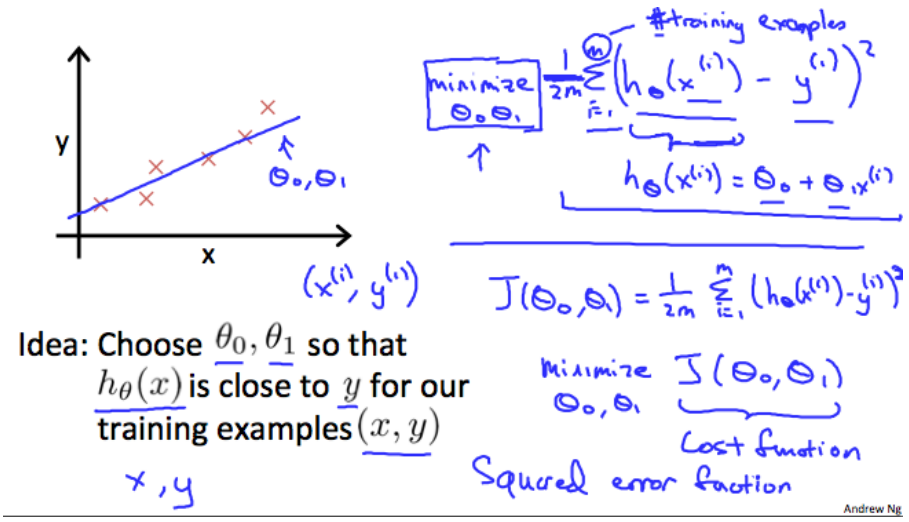

Figura 2: Função de custo

Função de Custo - Exemplo

Visualmente, podemos plotar nosso conjunto de dados de treinamento no plano x-y. Estamos tentando fazer uma linha reta (definida por $h_\theta(x)$) passar através desses pontos de dados dispersos.

Nosso objetivo é obter a melhor reta possível. A melhor reta possível será tal que a média das distâncias verticais ao quadrado dos pontos dispersos até a reta é mínimo. Idealmente, a reta deve passar por todos os pontos do nosso conjunto de dados de treinamento. Neste caso, o valor de $J(\theta_0, \theta_1)$ será zero. O exemplo abaixo mostra a situação ideal onde temos a função de custo igual a zero.

Quando $\theta_1 = 1$, obtemos uma inclinação de 1 tal que o modelo (reta) resultante passa sobre todos os pontos de dados. Da mesma maneira, quando $\theta_1 = 0.5$, vemos a distância vertical dos pontos até a reta aumentar.

Isto aumenta nosso custo $J(\theta_1)$ para 0.58. Plotando outros pontos, obtemos o seguinte gráfico para a função de custo:

Nosso objetivo é minimizar o erro, e portanto, a função de custo. Para que a função atinga o mínimo de erro, devemos escolher os parâmetros de $h_\theta(x)$ que minimizem essa função, ou seja, $\theta_1 = 1$.

Função de Custo - Exemplo 2

Uma gráfico de contorno é aquele contém muitas linhas de contorno. Uma linha de contorno de uma função de duas variáveis tem um valor constante em todos os pontos pertencentes à mesma linha. Um exemplo desse gráfico é o da direita abaixo.

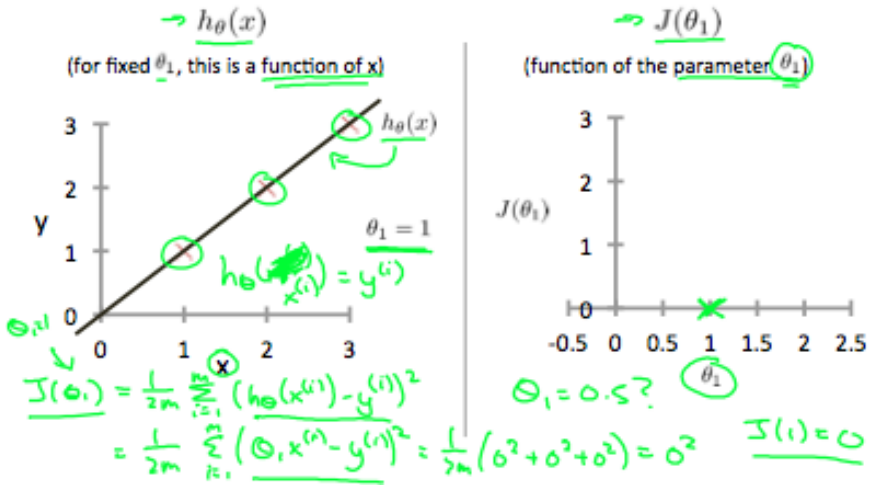

Figura 3: função de custo igual a zero

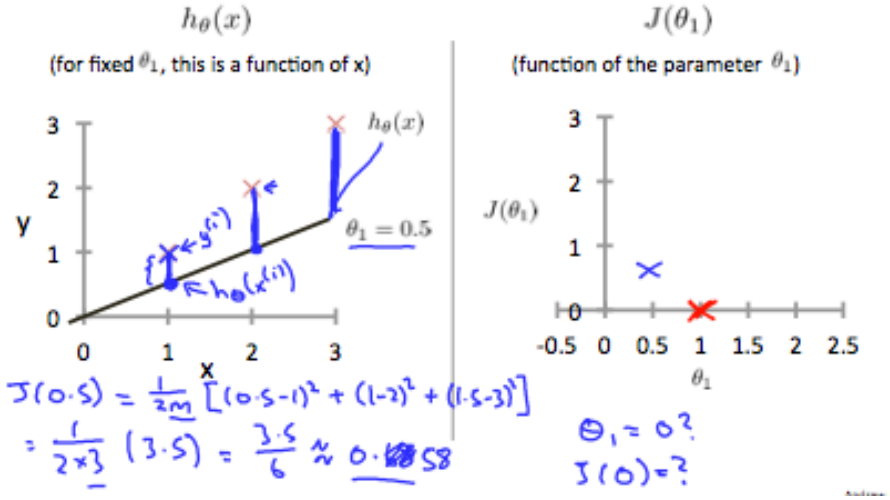

Figura 4: Distância vertical na função de custo.

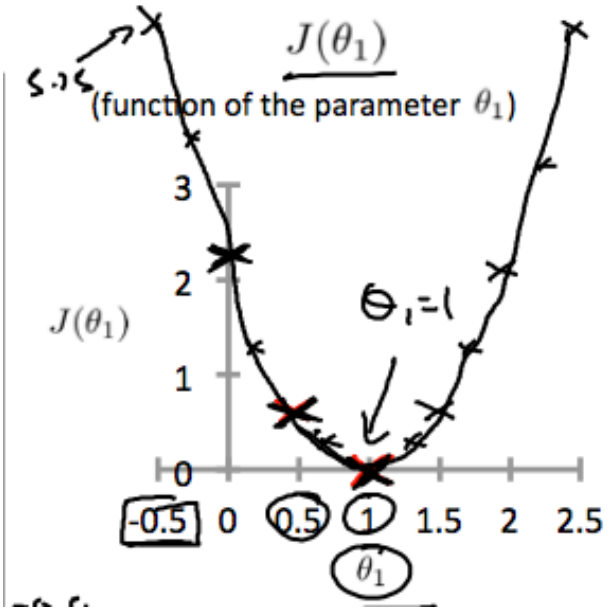

Figura 5:

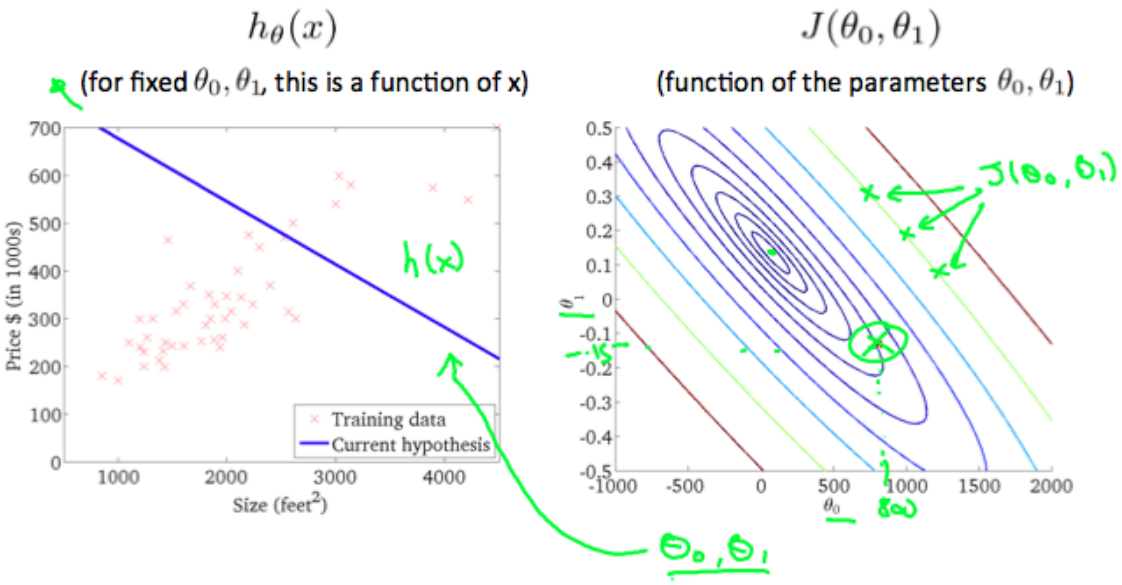

Figura 6:

Escolhendo qualquer cor e seguindo pela "elipse"dessa cor (linha de contorno), esperamos obter o mesmo valor da função de custo. Por exemplo, os três pontos verdes marcados na linha verde têm o mesmo valor de $J(\theta_0, \theta_1)$ e, portanto, estão na mesma linha. O X circulado marca uma configuração dos parâmetros da hipótese $h_\theta(x)$ que determina um valor específico para a função de custo, ou seja, dentre as infinitas hipóteses possíveis, o X marca uma delas: $\theta_0 = 800$ e $\theta_1 = -0.15$ (à esquerda). As curvas de contorno no gráfico acima exibem o nível de valor da função de custo $J(\theta_0, \theta_1)$ para uma infinidade de configurações θ_0, θ_1 da hipótese. O mínimo dessa função é encontrado dentro da menor "elipse". Escolhendo outra hipótese $\theta_0 = 360$ e $\theta_1 = 0$, temos:

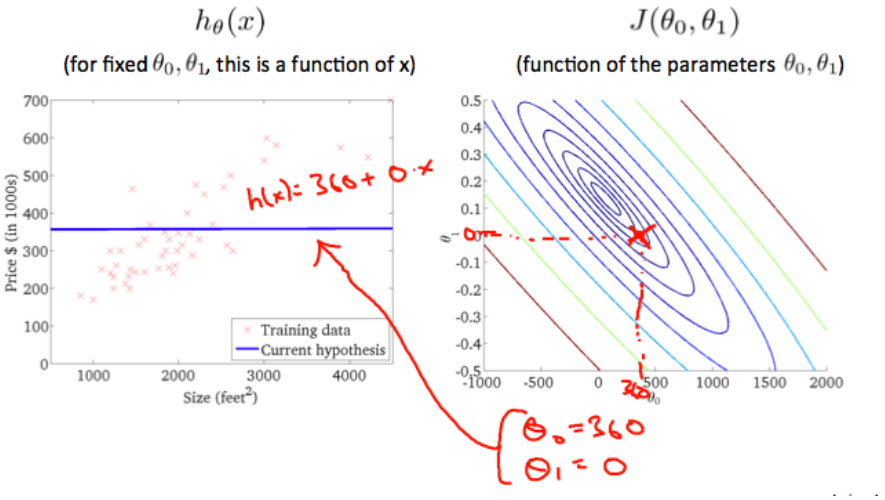

Figura 7:

Essa configuração faz com que $J(\theta_0, \theta_1)$ no gráfico de contorno se aproxime do centro, reduzindo o erro da função de custo. Agora, aumentando levemente a inclinação da nossa hipótese, obtemos um melhor ajuste (*fit*) aos dados.

O gráfico acima minimiza a função de custo tanto quanto possível e consequentemente, os resultados de θ_0, θ_1 ficam perto de 250 e 0.12, respetivamente. Plotando estes valores em nosso gráfico da direita, o ponto resultante parece estar no centro da menor "elipse".

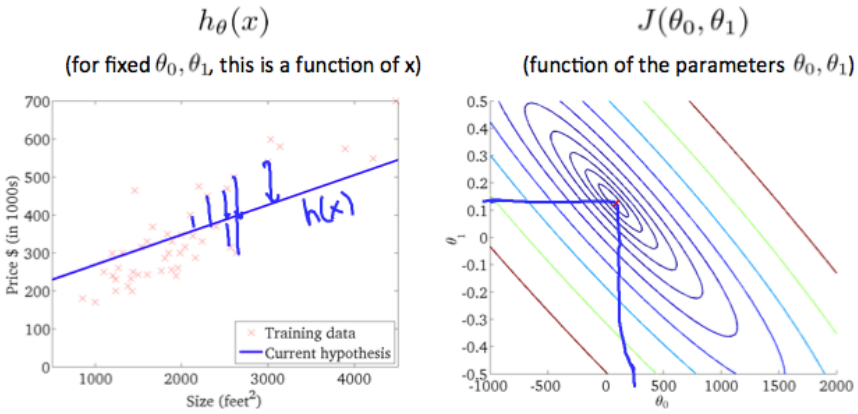

Figura 8:

Bibliografia

- Curso de Machine Learning de Andrew Ng.
- Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009. (`https://web.stanford.edu/~hastie/Papers/ESLII.pdf`)