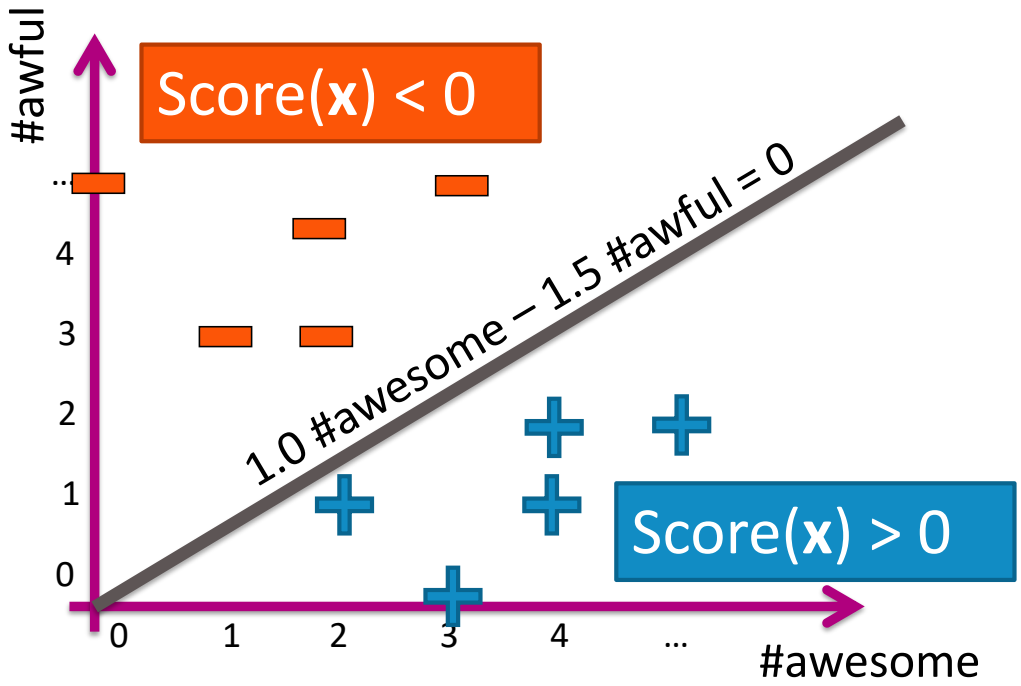
Review Classification

Linear classifier

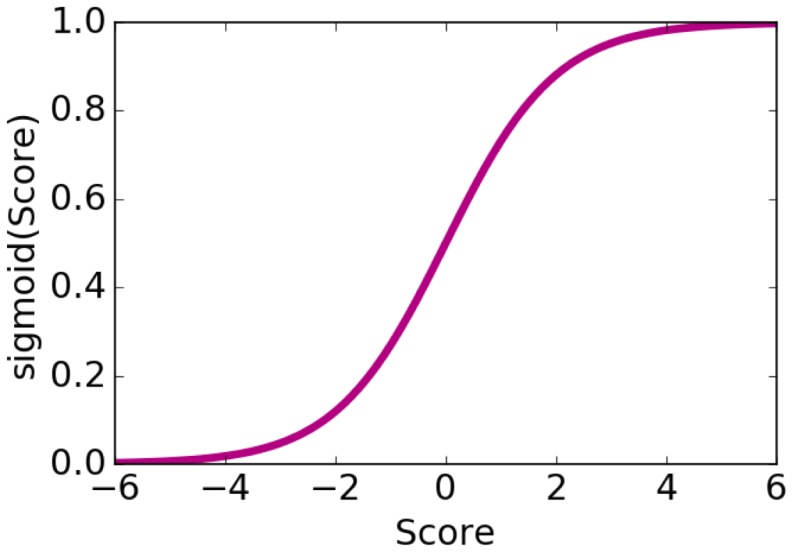
$$\text{Score}(x_i) = w_0\, h_0(x_i) + w_1\, h_1(x_i) + \ldots + w_D\, h_D(x_i)$$
$$= w^T h(x_i)$$



Logistic function (sigmoid, logit)

$$sigmoid(\text{Score}) = \frac{1}{1 + e^{-\text{Score}}}$$

Score	-∞	-2	0.0	+2	+∞
sigmoid(Score)					



Probabilistic Model

Distribution over sales prices

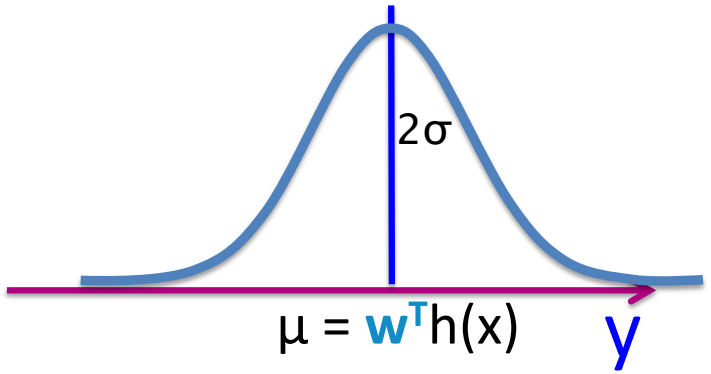
For houses with a given # sq.ft. (), what house prices $ are we likely to see?

Compare and contrast regression models

- Linear regression with Gaussian errors

  $y_i = \mathbf{w}^T h(x_i) + \varepsilon_i \quad \varepsilon_i \sim N(0,\sigma^2)$

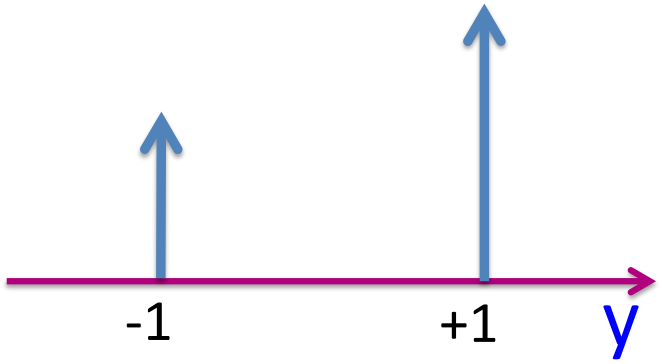
  ➔ $p(y|x,\mathbf{w}) = N(y; \mathbf{w}^T h(x), \sigma^2)$



Compare and contrast regression models

- Logistic regression

$$P(y|x,\mathbf{w}) = \begin{cases} \dfrac{1}{1 + e^{-\mathbf{w}^T h(x)}} & y = +1 \\ \dfrac{e^{-\mathbf{w}^T h(x)}}{1 + e^{-\mathbf{w}^T h(x)}} & y = -1 \end{cases}$$

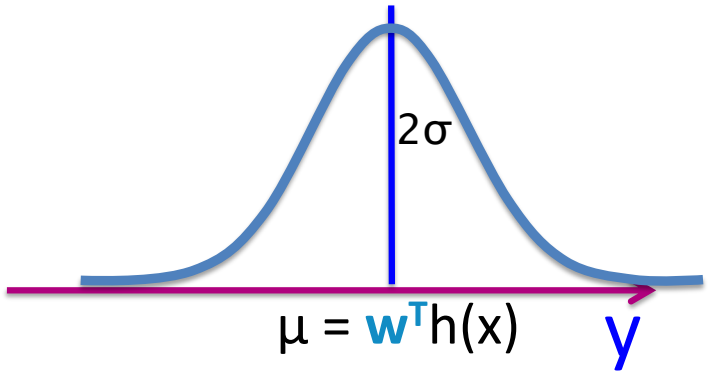


Compare and contrast regression models

- Linear regression with Gaussian errors

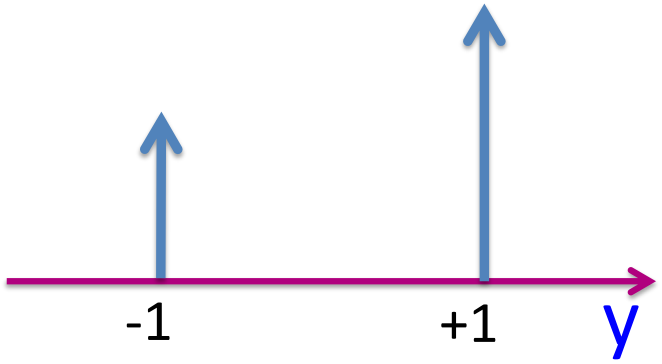
$y_i = \mathbf{w}^T h(x_i) + \varepsilon_i \quad \varepsilon_i \sim N(0,\sigma^2)$

➔ $p(y|x,\mathbf{w}) = N(y; \mathbf{w}^T h(x), \sigma^2)$

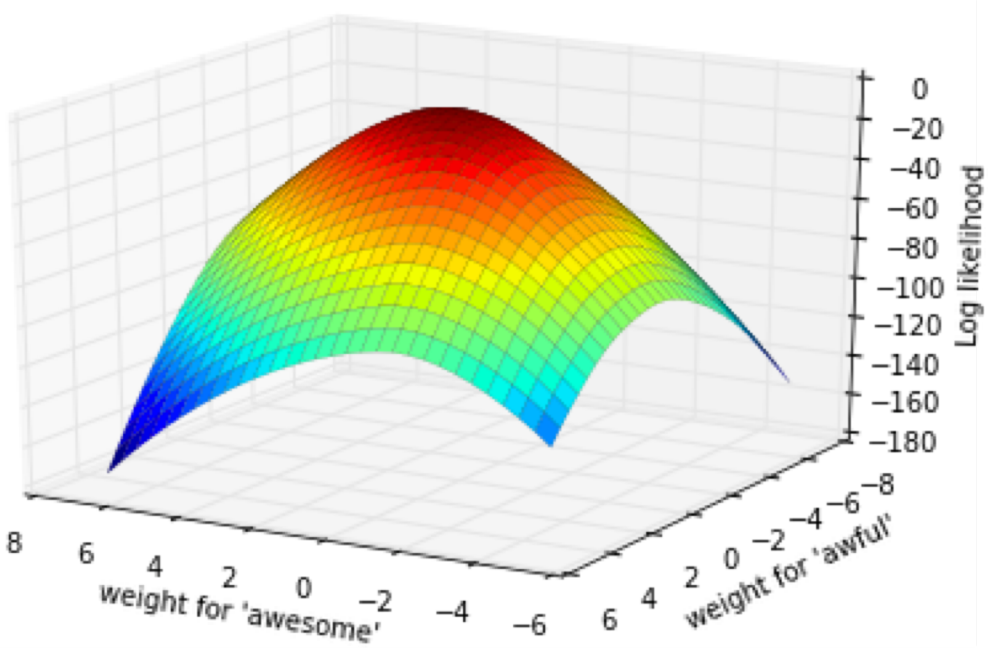


- Logistic regression

$$P(y|x,\mathbf{w}) = \begin{cases} \dfrac{1}{1 + e^{-\mathbf{w}^T h(x)}} & y = +1 \\ \dfrac{e^{-\mathbf{w}^T h(x)}}{1 + e^{-\mathbf{w}^T h(x)}} & y = -1 \end{cases}$$



Maximizing likelihood



Maximize function over all possible w_0, w_1, w_2

$$\max_{w_0,w_1,w_2} \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$\ell(w_0, w_1, w_2)$ is a function of 3 variables

No closed-form solution ➔ use gradient ascent

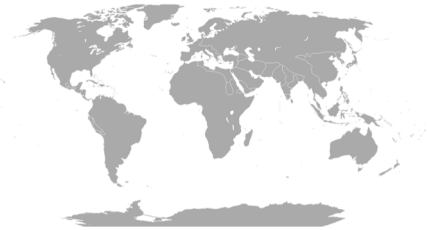
Encoding categorical inputs

Categorical inputs

- Numeric inputs:
  - #awesome, age, salary,…
  - Intuitive when multiplied by coefficient
    - e.g., 1.5 #awesome
- Categorical inputs:

Numeric value, but should be interpreted as category
(98195 not about 9x larger than 10005)

Gender
(Male, Female,…)



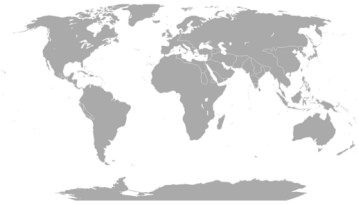
Country of birth
(Argentina, Brazil, USA,…)

Zipcode
(10005, 98195,…)

How do we multiply category by coefficient???
Must convert categorical inputs into numeric features

Encoding categories as numeric features

x =



Country of birth
(Argentina, Brazil, USA,...)

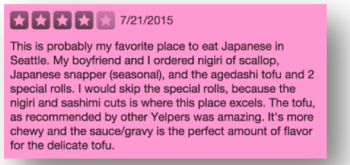
196 categories

1-hot encoding

x	$h_1(x)$	$h_2(x)$	...	$h_{195}(x)$	$h_{196}(x)$
Brazil					
Zimbabwe					

196 features

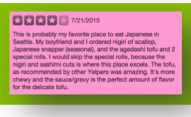
x =



Restaurant review
(Text data)

10,000 words in vocabulary

Bag of words

x	$h_1(x)$	$h_2(x)$	...	$h_{9999}(x)$	$h_{10000}(x)$
					

10,000 features