# introduction to regression

# continuous regression

- predict house price based on features (sq ft, # bathrooms, neighborhood school)
- forecast temperature in Indianapolis next week
- predict customer lifetime (years) they will stay w/ insurance company

# binary classification

- predict if a claim is fraudulent or not
- classify if image is of a cat or dog
- predict based on MRI if person has disease or not
- multi-class: hand written digit is 0-9 (10 classes…think atm check reader)

# generalized linear model (GLM)

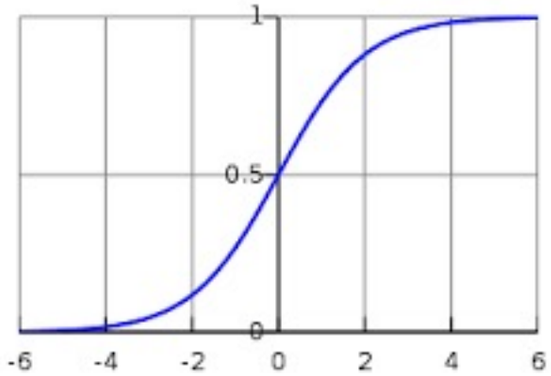
$$\theta = a_0 + a_1 x_1 + .... + a_n x_n$$

$$y \sim g(\theta)$$

- x are independent variables or “predictors”
- a are model coefficients (linear)
- theta is a parameter of the GLM (for regular linear regression g is the identity)

# logistic regression

$$\log(\frac{p}{1-p}) = a_0 + a_1 x_1 + .... + a_n x_n$$




$$p = \sigma(a_0 + a_1 x_1 + .... + a_n x_n)$$

$$y = \text{Bernoulli}(p)$$

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

# Gaussian (linear) regression

$$\mu = a_0 + a_1 x_1 + \dots$$

$$y = N(\mu, \sigma^2)$$

# maximum likelihood

- given a probability density and data, choose the parameters that maximize the probability of the data conditioned on the parameters

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \Pi_{i=1}^{N} f(x_i|\theta)$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{N} \log(f(x_i|\theta))$$

# maximum likelihood for Bernoulli

- Let y be 1 or 0 with probability p and 1-p
- What is the probability of observing 10110 given p?

# maximum likelihood for Bernoulli

- Let y be 1 or 0 with probability p and 1-p
- What is the probability of observing 10110 given p?

$$p(1-p)pp(1-p)$$

$$\sum_{i=1}^{5} p^{y_i}(1-p)^{1-y_i}$$

# maximum likelihood for logistic regression

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

$$p_i = \sigma(a_0 + a_1 x_1^i + .... + a_n x_n^i)$$

- can solve this approximately with a quasi-Newton method

# maximum likelihood for linear regression

$$L = \Pi_i \exp\left( -(y_i - X_i \vec{a})^2 / (2\sigma^2) \right)$$

$$logL = -\sum_i (y_i - X_i \vec{a})^2$$

- can solve this analytically (quadratic optimization)

# GLM in R

- In R use function glm with family set to binomial
- Run summary on trained model object to get variable p values and model goodness of fit
- Use predict function to apply model to new data

# Kaggle and common task framework

- we would like to mimic how a predictive model will perform in real life

- split data into three parts: training, test, validation

- you are given variables for all data, but only labels for the training dataset

- build a model to the train dataset, predict on test and validation

- you can only see performance on the test set until competition is over (to prevent over-fitting)

# metrics to evaluate models

- mean square error or  mean absolute error (continuous random variable)
- accuracy (% of samples correctly classified for a chosen threshold)
- precision (% of predicted positives that are actually positives)
- recall (% of positives falling in the predicted positive class)
- log-likelihood (probability of observing the data given the model)
- Many others that are application specific
- Often metrics are only meaningful when comparing competing models and considering the application
    - ex: log-likehood value has no intuitive meaning, but can rank models
    - ex: false positive rate in advertising can be high, it needs to be low in medical diagnostics