## Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

#### Term Explanations

P(A|B) = posterior probability of A given B

P(B|A) = likelihood of B given A

P(A) = prior probability of A

P(B) = marginal likelihood of B



### Standardization (Z-score)

$$z_i = \frac{x_i - \mu}{\sigma}$$

#### Term Explanations

 $z_i = \text{standardized score for data point } x_i$ 

 $x_i = \text{original data point } i$ 

 $\mu = \text{mean of the data}$ 

 $\sigma = \text{standard deviation of the data}$ 

# # /

### Mean Square Error (MSE)

MSE = 
$$\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

#### Term Explanations

MSE = Mean Square Error  $y_i = \text{actual value for data point } i$   $\hat{y}_i = \text{predicted value for data point } i$  m = number of data points

## **Multi-class Cross Entropy Loss**

$$L = -\sum_{i=1}^{m} \sum_{k=1}^{K} y_{ik} \log(\hat{p}_{ik})$$

#### Term Explanations

L = cross entropy loss

 $y_{ik} = 1$  if true label of i is k else 0

 $\hat{p}_{ik}$  = predicted probability of *i* being in class *k* 

m = number of data points

K = number of classes



#### **Softmax Function**

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

#### Term Explanations

 $\sigma(z_i) = \text{softmax score for class } i$  $e^{z_i} = \text{exponentiated score for class } i$ 

 $\sum_{i=1}^{K} e^{z_i} = \text{sum of exponentiated scores for all classes}$ 

## **Gradient Descent Update Rule**

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} J(\mathbf{w})$$

#### Term Explanations

 $\mathbf{w} = \text{weight vector}$ 

 $\eta = \text{learning rate}$ 

 $\nabla_{\mathbf{w}} J(\mathbf{w}) = \text{gradient of the loss function } J \text{ with respect to } \mathbf{w}$ 



# Linear Regression Normal Equations (MATRIX)

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

#### Term Explanations

 $\mathbf{w} = \text{weight vector}$ 

X = data matrix (features)

 $\mathbf{X}^T$  = transpose of data matrix

y = vector of true values

 $(\mathbf{X}^T\mathbf{X})^{-1}$  = inverse of matrix product

## Logistic Regression

$$\hat{\mathbf{y}} = \sigma(\mathbf{X}\mathbf{w} + \mathbf{b})$$

$$J(\mathbf{w}, \mathbf{b}) = -\frac{1}{m} \left( \mathbf{y}^T \log(\hat{\mathbf{y}}) + (1 - \mathbf{y})^T \log(1 - \hat{\mathbf{y}}) \right)$$

#### Term Explanations

 $\hat{\mathbf{y}}$  = vector of predicted probabilities

$$\sigma(z) = \text{sigmoid function } \sigma(z) = \frac{1}{1 + e^{-z}}$$

X = data matrix (features)

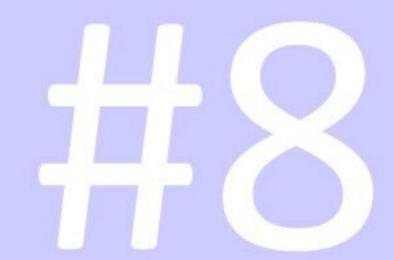
 $\mathbf{w} = \text{weight vector}$ 

 $\mathbf{b} = \text{bias vector}$ 

 $J(\mathbf{w}, \mathbf{b}) = \cos t \text{ function}$ 

y = vector of true labels

m = number of data points



## K-means Clustering Objective

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} \mathbf{1}_{\{c_i = k\}} ||x_i - \mu_k||^2$$

#### Term Explanations

J = total within-cluster sum of squares

 $\mathbf{1}_{\{c_i=k\}} = \text{indicator function for data point } x_i \text{ in cluster } k$ 

 $x_i = \text{data point } i$ 

 $\mu_k = \text{centroid of cluster } k$ 



## **Principal Component Analysis**

$$\Sigma = \frac{1}{m} X^T X$$

$$w = \operatorname{argmax}_{w} \left( w^{T} \Sigma w \text{ subject to } ||w|| = 1 \right)$$

#### Term Explanations

 $\Sigma$  = covariance matrix

X = data matrix

m = number of data points

w = principal component vector