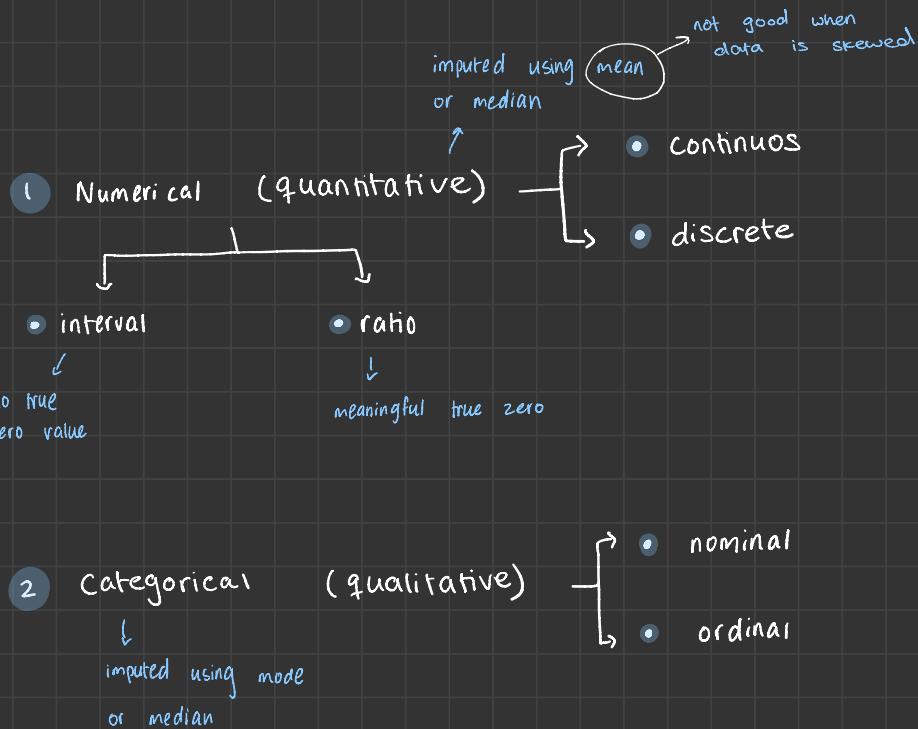


week 1 and week 2 → useless

WEEK 3: Data quality exploration

— types of features —



— Correlations —

1 numerical VS numerical :

• Pearson → $(-1, 1)$

- ✓ continuos variables
- ✓ normal distributions
- measures linearity

• Spearman → $(-1, 1)$

- ✓ 2 continuos or 2 ordinal categoricon
- ✓ normal or skewed
- ↳ measures strength and direction, regardless of linearity

2 numerical VS categorical

• Point - biserial

pearson's but
for numerical
VS binary

↓
(-1, 1)
↓
-ve or
+ve relation

• ANOVA

good for
multiple categorises
(more than binary)

3 categorical VS categorical

• Chi-square → tests for independence

— data preparation —

1 categorical attributes

label / ordinal encoding	nominal / one-hot encoding
1 st class → 0	man 0 0 1
2 nd class → 1	woman 0 1 0
⋮	child 1 0 0
n th class → n-1	

2 numerical attributes

• min-max scaling

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

range: [0, 1]

• Standard scaler

$$z = \frac{x - \mu}{\sigma}$$

$x \sim N(0, 1)$

mean
↑
standard deviation

correlation examples

1 Pearson's correlation coefficient

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \rightarrow \begin{matrix} \text{covariance} \\ \rightarrow \text{standard deviation} \end{matrix}$$

$$\text{cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$= \frac{13.5 + 0.5 + 1.5 + 10.5}{4}$$

$$\text{cov}(X, Y) = 6.5$$

$$\sigma_X = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} = \sqrt{\frac{20}{4}} = \sqrt{5}$$

X	Y
2	3
4	7
6	9
8	11

$$\bar{X} = 5$$

$$\bar{Y} = 7.5$$

$$\sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{35}{4}} = \frac{\sqrt{35}}{2}$$

$$r = \frac{6 \cdot 5}{(\sqrt{5})(\frac{\sqrt{35}}{2})} = \frac{13\sqrt{7}}{35} = 0.983$$

\Rightarrow strong positive linear relationship

2 Spearman's correlation coefficient

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \rightarrow \text{Rank}(x_i) - \text{Rank}(y_i)$$

	X	Y	
1	97	195	1
2	99	195	2] $\rightarrow 1.5$
3	101	210	5
4	103	205	3] $\rightarrow 3.5$
5	105	205	4

	X	Y
3	101	210
2	99	195
4	103	205
1	97	195
5	105	205

$$\sum d_i^2 = 0.25 + 0.25 + 4 + 0.25 + 2.25 = 7$$

$$\rho = 1 - \frac{6(7)}{5(5^2 - 1)} = 1 - \frac{42}{120} = 0.65$$

The Spearman rank correlation coefficient ρ is 0.65, indicating a moderate positive monotonic relationship between X and Y . As the values of X increase, the values of Y tend to increase, but not in a perfectly linear fashion.

3 Point Biserial

means of class 1 & 0

$$\rho = \frac{M_1 - M_0}{\sigma} \times \sqrt{P_1 P_0}$$

standard deviation of Y

Gender(X)	Test Score(Y)
Male (1)	85
Female (0)	78
Male (1)	90
Female (0)	75
Male (1)	88
Female (0)	80

$$M_1 = \frac{85 + 90 + 88}{3} = \frac{263}{3} = 87.67$$

$$M_0 = \frac{78 + 75 + 80}{3} = \frac{233}{3} = 77.67$$

$$P_1 = \frac{3}{6} = 0.5$$

$$P_0 = \frac{3}{6} = 0.5$$

$$\sigma = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$$

$$\bar{y} = \frac{246}{3} = 82.67$$

$$\sigma = \sqrt{\frac{175.3334}{6}} = 5.406$$

$$\rho = \frac{87.67 - 77.67}{5.406} \times \sqrt{0.5 \times 0.5} = 0.9249$$

The point biserial correlation coefficient $\rho = 0.925$, indicating a strong positive relationship between gender and test scores. Males tend to have higher test scores than females in this dataset.

4 ANOVA

$$F\text{-statistic} = \frac{MSB}{MSW}$$

↑ mean sum of square between groups
↓ " " within groups

$$MSB = \frac{\sum n_i (\bar{x}_i - \bar{x})^2}{df_B}$$

↑ sum of squares between
↑ # of class i
↑ mean of class i
↑ grand mean
↓ degrees of freedom

$$MSW = \frac{\sum \sum_i (x_{ij} - \bar{x}_j)^2}{\# \text{ of observations} - \# \text{ of groups}}$$

↑ j ∈ classes i ∈ each class
↑ observation in class i
↑ class mean

$$\text{grand mean} = \frac{760}{9} = 84.44$$

$$\text{class A mean} = \frac{263}{3} = 87.67$$

$$\text{class B mean} = \frac{153}{2} = 76.5$$

$$\text{class C mean} = \frac{344}{4} = 86$$

Test Scores	Teaching Method
85	A
90	
88	
75	B
78	
82	C
85	
87	
90	

$$df_B = 3-1 = 2 \quad df_W = 9-3 = 6$$

$$MSB = \frac{3(87.67 - 84.44)^2 + \dots}{2} = \frac{167.1203}{2} = 83.56$$

$$SSW_A = (85 - 87.67)^2 + \dots = 12.67$$

$$SSW_B = 4.5$$

$$SSW_C = 34$$

$$MSW = \frac{12.67 + 4.5 + 34}{6} = 8.53$$

$$F\text{-statistic} = \frac{83.56}{8.53} = 9.796$$

Comparing F-statistic w/ critical value from table
at p-value = 0.05 or 0.01 for $df(x, y)$
if $(F\text{-statistic} \geqslant \text{critical}) \Rightarrow \text{significant}$ at $\alpha = 0.05 | 0.01$

$$df(2, 6)$$

$$\alpha = 0.05 \rightarrow \text{critical} = 5.14 \leq 9.796 \Rightarrow \text{significant}$$

$$F = 9.796$$

At $\alpha = 0.05$, the critical value for $df_B = 2$ and $df_W = 6$ is 5.14.

Since $F = 9.79$ is greater than the critical value, this indicates that there is a statistically significant difference in test scores between the teaching methods.

5 chi-square independence test

$$\chi^2 = \sum_{\substack{\text{across the} \\ \text{combination of} \\ \text{two classes}}} \frac{(O - E)^2}{E}$$

observed frequency ↑ expected frequency →

$$= \frac{\text{total class A} \times \text{total class B}}{\text{grand total}}$$

$$df = (r-1)(c-1)$$

↑
of classes
in 2nd group

↑
of classes
in 1st group

	Method A	Method B	Method C	Total
Male	30	10	20	60
Female	10	30	50	90
Total	40	40	70	150

$$\text{Male / method A expected frequency} = \frac{60 \times 40}{150} = 16$$

table with all expected frequencies :

	method A	method B	method C
male	16	16	28
female	24	24	42

$$\text{male / method A } \chi^2 = \frac{(30 - 16)^2}{16} = 12.25$$

table with χ^2 values

	method A	method B	method C
male	12.25	2.25	2.286
female	8.167	1.5	1.524

$$\chi^2 = \sum x_i^2 = 27.977$$

$$df = (2-1)(3-1) = 1 \times 2 = 2$$

comparing with critical value from table at $\alpha = 0.05$

if ($\chi^2 >$ critical) \rightarrow significant

$$\left. \begin{array}{l} \chi^2 = 27.977 \\ \alpha = 0.05 \\ df = 2 \end{array} \right\} \text{critical} = 5.991 \leq 27.977 \Rightarrow \text{significant}$$

The critical value for $df = 2$ and $\alpha = 0.05$ is 5.99. Since $\chi^2 = 27.98$, this indicates there is statistically significant correlation between gender and teaching methods.

Week 4: linear regression

— univariate linear regression —

- hypothesis: $h_w(x) = w_0 + w_1 x$

- cost function: $J(w) = \frac{1}{2m} \sum_{i=1}^m (h_w(x_i) - y_i)^2$
mean sum of square error $\xrightarrow{m \rightarrow \text{sample size}}$
 $\sqrt{\quad}$ prediction \downarrow true label

goal : $\min_w J(w)$ learning rate

repeat until convergence : $w_j := w_j - \alpha \frac{d J(w)}{d w_j}$
derivative of $J(w)$ w.r.t. w_j

• gradient descent

Let $J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^m (h_w(x_i) - y_i)^2$ s.t.

$h_w(x) = w_0 + w_1 x$, then:

$$J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^m (w_0 + w_1 x_i - y_i)^2$$

$$\frac{\partial}{\partial w_0} J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^m \cancel{x}(w_0 + w_1 x_i - y_i) \cdot 1 \quad \begin{matrix} \downarrow \\ \frac{\partial}{\partial w_0} h_w(x_i) \end{matrix}$$

$$\frac{\partial}{\partial w_1} J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^m \cancel{x}(w_0 + w_1 x_i - y_i) \cdot x_i \quad \begin{matrix} \uparrow \\ \frac{\partial}{\partial w_1} h_w(x_i) \end{matrix}$$

Standard gradient descent (SGD) update algorithm:

$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_w(x_i) - y_i)$$

$$w_1 := w_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_w(x_i) - y_i) \cdot x_i$$

Multivariate linear regression

$y \in \mathbb{R}^m$ $w \in \mathbb{R}^m$ for m samples
↑ ↑
• $h_w(x) = w_0 + w_1 x_1 + \dots + w_n x_n$ $x \in \mathbb{R}^{n \times n}$ ↑ number of features

• SGD: $w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_w(x) - y_i) \cdot x_{ij}$
↓
sample i
feature j

• convergence: when $\Delta J(w) \leq 10^{-3}$ after an iteration

• vector representation:

$$w^\top = [w_0 \ w_1 \ \dots \ w_n] \quad w \in \mathbb{R}^{n+1}$$

$$x = \begin{bmatrix} x_0^{(1)} & \dots & x_0^{(m)} \\ x_1^{(1)} & \dots & : \\ \vdots & & \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix} \quad \begin{array}{l} \text{all 1's for intercept} \\ (n+1) \times m \rightarrow m \text{ samples} \end{array}$$

$$\Rightarrow h_w(x) = w^\top x \quad h_w(x) \in \mathbb{R}^{m \text{ predictions}}$$

$$\Rightarrow J(\omega) = \frac{1}{2m} \sum_{i=1}^m (\omega^T x^{(i)} - y^{(i)})^2$$

$$\text{or } J(\omega) = \frac{1}{2m} (\omega^T x - y)^T (\omega^T x - y)$$

$$\Rightarrow w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^m (\omega^T x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

normal equation -

analytically find w_j when $\frac{\partial}{\partial w_j} J(\omega) = 0 \quad \forall j \in [0, n]$

Let $\omega \in \mathbb{R}^{(n+1) \times 1}$, $X \in \mathbb{R}^{m \times (n+1)}$, $y \in \mathbb{R}^{m \times 1}$, then:

changed from single sample
example but whatever

$$J(\omega) = \frac{1}{2m} (\underbrace{X\omega}_{m \times 1} - y)^T (X\omega - y)$$

$$X\omega \cdot X\omega = \underbrace{\omega^T X^T X \omega}_{l \times (n+1) \times (n+1) \times l} = \underbrace{\omega^T y}_{m \times 1}$$

scalar

$$J(\omega) = \frac{1}{2m} (\omega^T X^T X \omega - 2\omega^T X^T y - y^T y)$$

$$X\omega \cdot y = \underbrace{\omega^T X^T y}_{l \times (n+1) \times m \times 1}$$

$$\frac{\partial}{\partial \omega} J(\omega) = \frac{1}{2m} (2X^T X \omega - 2X^T y)$$

$$\cancel{\frac{1}{m}} (X^T X \omega - X^T y) = 0$$

$$X^T X \omega = X^T y$$

$$\boxed{\omega = (X^T X)^{-1} X^T y}$$

\downarrow
 $n \times n \rightarrow n \times n$
 $n \times n \rightarrow n \times 1$

→ substitute & find
optimal ω vector

Ordinary least squares

Special case of normal equation with one feature only:
 $n = 1$

$$\Rightarrow w \in \mathbb{R}^{2 \times 1}, x \in \mathbb{R}^{2 \times M}, y \in \mathbb{R}^{M \times 1}$$

$$w_1 = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

OLS example

$$\bar{x} = 2.5 \quad \bar{y} = 4.25$$

$$w_1 = \frac{\sum (x_i - 2.5)(y_i - 4.25)}{\sum (x_i - 2.5)^2}$$

x	y
1	2
2	3
3	5
4	7

$$w_1 = \frac{3.375 + 0.625 + 0.375 + 4.125}{2.25 + 0.25 + 0.25 + 2.25}$$

$$= \frac{8.5}{5} = 1.7$$

$$w_0 = 4.25 - (1.7)(2.5) = 0$$

$$h_w(x) = 1.7x$$

evaluation metrics

$$RMSE \underset{\text{(root mean square error)}}{=} \sqrt{\frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2}$$

$\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$ $\overbrace{\phantom{\sum_{i=1}^n}}$ SS_{res}

$$SS_{tot} = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

$y^{(i)}$ true label \bar{y} mean of true label

$$SS_{reg} = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2$$

$\hat{y}^{(i)}$ prediction

$$SS_{res} = \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \rightarrow \text{assesses how well the model explains variance in the data}$$

$$\text{OR } R^2 = \frac{SS_{reg}}{SS_{tot}}$$

$$\text{Adjusted } R^2 = \frac{1 - (1 - R^2)(m - 1)}{m - n - 1} \rightarrow \text{decreases with the addition of irrelevant features}$$

$m - n - 1$
 $\curvearrowleft \quad \downarrow$
sample size # of features

Week 5: Logistic regression

$$h_w(x) = \sigma(w^T x)$$

↑
sigmoid

s.t. $0 \leq h_w(x) \leq 1$

↓
probability

$$= \frac{1}{1 + e^{-w^T x}}$$

- prediction threshold = 0.5
- class 1 if $h_w(x) \geq 0.5$
- class 0 if $h_w(x) < 0.5$

$$\text{cost}(h_w(x), y) = \begin{cases} -\log(h_w(x)) & \text{if } y=1 \\ -\log(1-h_w(x)) & \text{if } y=0 \end{cases}$$

approaches 0/1, never exact

$-\log(1) = 0$
 $-\log(0) \rightarrow \infty$

$-\log(1-\alpha) = 0$
 $-\log(1-\alpha) \rightarrow \infty$

$$\text{Cost}(h_w(x), y) = y \cdot -\log(h_w(x)) + (1-y) \cdot -\log(1-y)$$

↓ "activates" the corresponding cost when $y=1$

↓ "activates" the corresponding cost when $y=0$

$$J(w) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_w(x^{(i)}), y^{(i)})$$

$$J(w) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_w(x^{(i)})) + (1-y^{(i)}) \log(1-h_w(x^{(i)}))$$

goal: minimize $J(w)$

$$\text{update rule: } w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w)$$

deriving $\frac{\partial}{\partial w_j} J(\omega)$

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_\omega(x^{(i)})) + (1-y^{(i)}) \log(1-h_\omega(x^{(i)}))$$

we know that $\frac{d}{dx} \log(f(x)) = \frac{f'(x)}{f(x)}$. Then:

$$\frac{\partial}{\partial w_j} J(\omega) = -\frac{1}{m} \sum_{i=1}^m \frac{y^{(i)}}{h_\omega(x^{(i)})} \cdot \boxed{\frac{\partial}{\partial w_j} h_\omega(x^{(i)})}$$

$$+ \frac{1-y^{(i)}}{1-h_\omega(x^{(i)})} \cdot \boxed{\frac{\partial}{\partial w_j} 1-h_\omega(x^{(i)})}$$

evaluating $\frac{\partial}{\partial w_j} h_\omega(x)$:

$$h_\omega(x) = \sigma(\omega^\top x)$$

$$\frac{d}{dx} \sigma(f(x)) = \sigma(f(x))(1-\sigma(f(x))) = f'(x)$$

$$\Rightarrow \frac{d}{\partial w_j} h_\omega(x) = \sigma(\omega^\top x)(1-\sigma(\omega^\top x)) \cdot \frac{\partial}{\partial w_j} \omega^\top x$$

$$\frac{d}{\partial w_j} h_\omega(x) = \sigma(\omega^\top x)(1-\sigma(\omega^\top x)) \cdot x_j \xrightarrow{\text{feature } j}$$

$$\hookrightarrow x_0 = 1$$

because intercept w_0

similarly :

$$\frac{\partial}{\partial w_j} (1 - h_w(x)) = - \frac{\partial}{\partial w_j} h_w(x) = - \sigma(w^T x) (1 - \sigma(w^T x)) \cdot x_j^{(i)}$$

substituting back into $\frac{\partial J(\omega)}{\partial w_j}$:

$$\begin{aligned} \frac{\partial J(\omega)}{\partial w_j} &= -\frac{1}{m} \sum_{i=1}^m \frac{y^{(i)}}{\sigma(w^T x)} \cdot \sigma(w^T x) (1 - \sigma(w^T x)) \cdot x_j^{(i)} \\ &\quad + \frac{1 - y^{(i)}}{1 - \sigma(w^T x)} \cdot -\sigma(w^T x) (1 - \sigma(w^T x)) \cdot x_j^{(i)} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial w_j} J(\omega) &= -\frac{1}{m} \sum_{i=1}^m y^{(i)} (1 - \sigma(w^T x)) x_j^{(i)} - (1 - y^{(i)}) \sigma(w^T x) x_j^{(i)} \\ &\quad y^{(i)} x_j^{(i)} - y^{(i)} x_j^{(i)} \sigma(w^T x) - x_j^{(i)} \sigma(w^T x) + y^{(i)} x_j^{(i)} \sigma(w^T x) \\ &= -\frac{1}{m} \sum_{i=1}^m y^{(i)} x_j^{(i)} - x_j^{(i)} \sigma(w^T x) \stackrel{h_w(w^T x)}{\uparrow} \\ &= -\frac{1}{m} \sum_{i=1}^m (y^{(i)} - h_w(x^{(i)})) x_j^{(i)} \end{aligned}$$

$$\boxed{\frac{\partial}{\partial w_j} J(\omega) = \frac{1}{m} \sum_{i=1}^m (h_w(x^{(i)}) - y^{(i)}) x_j^{(i)}}$$

all of this is for binary classification, not multiclass

evaluation metrics

		Predicted	
		class 0	class 1
Actual	class 0	TP FN	FP TP
	class 1		

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

after targeting tue, how many we correct / "precise" ?

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

out of all the actual tues, how many were correct / "recalled" ?

similar to accuracy for the only

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} + \text{Recall}}{\text{Precision} \cdot \text{Recall}}$$

harmonic mean of precision & recall

In a logistic regression model, the interpretation of coefficients differs from linear regression due to the nature of the target variable, which is binary (0 or 1). Instead of predicting a continuous outcome, logistic regression predicts the **log-odds** of target variable being 1.

$$\mathbf{w}^T \mathbf{x} = -1 + 0.5x_1 - 0.8x_2$$

- When both x_1 and x_2 are zero, the log-odds of $y = 1$ are -1 . The odds are $e^{-1} \approx 0.37$, meaning the odds of $y = 1$ are 0.37 to 1. $P(y=1) = \frac{0.37}{1+0.37} = 0.27$.
- For a one-unit increase in x_1 , the log-odds of $y = 1$ increase by 0.5, or equivalently, the odds of $y = 1$ increase by a factor of $e^{0.5} \approx 1.65$, a 65% increase.
- For a one-unit increase in x_2 , the log-odds of $y = 1$ decrease by 0.8, or equivalently, the odds of $y = 1$ decrease by a factor of $e^{-0.8} \approx 0.45$, a 55% decrease.

week 5: regularization

to address overfitting:

- 1 reduce # of features : feature selection & dimension reduction
- 2 regularization : reduce magnitude of $w_j \in w$, makes model worse to avoid overfitting

ridge (L2) regression

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n w_j^2 \right]$$

↓
cost function

regularization parameter

gradient descent :

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w)$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{1}{2m} \left[2 \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda \underbrace{\frac{\partial}{\partial w_j} \sum_{j=1}^n w_j^2}_{\downarrow} \right]$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{1}{2m} \left[2 \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda \underbrace{2 w_j}_{\downarrow} \right]$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{1}{m} \left[\sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda w_j \right]$$

$$w_j := w_j - \frac{\alpha}{m} \left[\sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda w_j \right]$$

$$w_j := w_j - \frac{\alpha}{m} \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) - \frac{\lambda \alpha}{m} w_j$$

$$w_j := w_j \left(1 - \frac{\lambda \alpha}{m} \right) - \frac{\alpha}{m} \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)})$$

normal equation

ridge (L2) regularization

$$w = (X^T X + \lambda \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix})^{-1} X^T y$$

only $\downarrow \lambda$ to w_j s.t. $j > 1$

slightly inc. the inverse \Rightarrow dec. magnitude of w

$$(n+1) \times (n+1) \quad I_{(n+1)}$$

$X = \begin{bmatrix} 2 & 5 & -3 \\ 7 & 1 & 8 \\ 4 & -2 & 6 \end{bmatrix}$	$y = \begin{bmatrix} 3 \\ -1 \\ 7 \end{bmatrix}$	$\lambda = 0.7$
---	--	-----------------

$$I_{4 \times 4}$$

$$m \times (n+1) \quad 3 \times 3$$

$$3 \times 1$$

$$\left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 7 & 4 \\ 5 & 1 & -2 \\ -3 & 8 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 5 & -3 \\ 1 & 7 & 1 & 8 \\ 1 & 4 & -2 & 6 \end{bmatrix} + 0.7 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 4 \\ 5 & 1 & -2 \\ -3 & 8 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 7 \end{bmatrix} = w \in \mathbb{R}^{4 \times 1}$$

lasso (l1) regression

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n |w_j| \right]$$

↓ cost function ↓ regularization parameter ↓ absolute value, not w_j^2

gradient descent :

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w)$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{1}{2m} \left[2 \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda \underbrace{\frac{\partial}{\partial w_j} \sum_{j=1}^n |w_j|}_{\downarrow} \right]$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{1}{2m} \left[2 \sum_{i=1}^n (h_w(x^{(i)}) - y^{(i)}) + \lambda \cdot \underbrace{\text{sign}(w_j)}_{\downarrow} \right]$$

$$\text{sign}(w_j) = \begin{cases} +1 & \text{if } w_j > 0 \\ -1 & \text{if } w_j < 0 \\ 0 & \text{if } w_j = 0 \end{cases}$$

\downarrow
 pulls w_j closer to 0
 during update \rightarrow good for feature selection

$$\frac{\partial}{\partial w_j} J(\omega) = \frac{1}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)}) + \frac{\lambda}{2m} \cdot \text{sign}(w_j)$$

$$w_j := w_j - \frac{\lambda}{2m} \text{sign}(w_j) - \frac{\alpha}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)})$$

logistic regression

ridge (L2) regularization

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_\omega(x^{(i)})) + (1-y^{(i)}) \log(1-h_\omega(x^{(i)}))$$

$$+ \frac{\lambda}{2m} \sum_{j=1}^n w_j^2$$

$$\frac{\partial}{\partial w_j} J(\omega) = \frac{1}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$$

$$w_j := w_j (1 - \frac{\lambda}{m}) - \frac{\alpha}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

logistic regression

lasso (L1) regularization

$$J(\omega) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_\omega(x^{(i)})) + (1-y^{(i)}) \log(1-h_\omega(x^{(i)}))$$

$$+ \frac{\lambda}{2m} \sum_{j=1}^n |w_j|$$

$$\frac{\partial}{\partial w_j} J(\omega) = \frac{1}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{2m} \cdot \text{sign}(w_j)$$

$$w_j := w_j - \frac{\alpha \lambda}{2m} \cdot \text{sign}(w_j) - \frac{\alpha}{m} \sum_{i=1}^m (h_\omega(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

WEEK 6: decision trees

$$\text{entropy } H(p_i) = - \sum_{i=1}^n p_i \log_2(p_i)$$

↓
 measure of
 impurity

n ↑ # of classes

↴ fraction of samples
 in class i

$H(p_i) = 0 \rightarrow 100\% \text{ pure}$ or or 100% from 1 class

$H(p_i) = 1 \rightarrow 100\% \text{ impure}$ 50% between 2 diff. classes

$$\text{entropy} = H(p_i) = -p_i \log_2(p_i) - p_o \log_2(p_o) \rightarrow \text{binary classification}$$

↑ p_i ↑ p_o

$$\text{information gain} = H(p_i^{\text{root}}) - \left[w^{\text{left}} H(p_i^{\text{left}}) + w^{\text{right}} H(p_i^{\text{right}}) \right]$$

↓
 entropy at root ↴ # of samples in left node

1 calculate information gain for all possible splits

2 choose highest information gain

3 Split & repeat until:

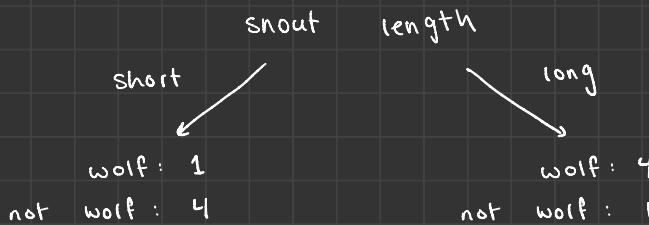
- $H(p_i) = 0$ for a node
- max. depth reached
- info. gain < threshold H splits
- # of samples in node < threshold

10 samples

Snout Length	Face Shape	Tail Bushiness	Wolf
Long	Broad	Less	1
Short	Narrow	Less	1
Short	Broad	More	0
Long	Narrow	Less	0
Long	Broad	Less	1
Long	Broad	More	1
Short	Narrow	More	0
Long	Broad	More	1
Short	Broad	More	0
Short	Broad	More	0

root node

$$H(p_{\text{root}}) = - \frac{5}{10} \log_2 \left(\frac{5}{10} \right) - \frac{5}{10} \log_2 \left(\frac{5}{10} \right) = 1$$



$$P_1 = \frac{1}{5}$$

$$w = \frac{5}{10}$$

$$P_1 = \frac{4}{5}$$

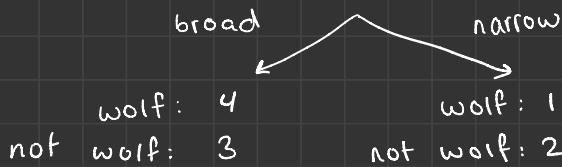
$$w = \frac{5}{10}$$

$$H(P_1) = - \frac{1}{5} \log_2 \left(\frac{1}{5} \right) - \frac{1}{5} \log_2 \left(\frac{1}{5} \right) = 0.722 \quad H(P_1) = 0.722$$

$$\begin{aligned} \text{info. gain} &= H(p_{\text{root}}) - \left[w^{\text{left}} H(p_{\text{left}}) + w^{\text{right}} H(p_{\text{right}}) \right] \\ &= 1 - \left[\frac{5}{10} (0.722) + \frac{5}{10} (0.722) \right] \end{aligned}$$

$$= 0.278$$

face shape



$$P_1 = \frac{4}{7}$$

$$\omega = \frac{7}{10}$$

$$H\left(\frac{4}{7}\right) = 0.985$$

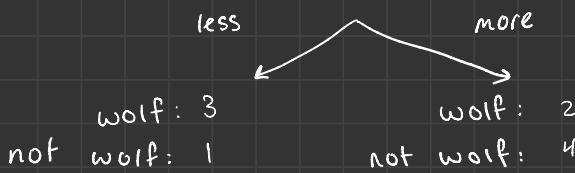
$$P_1 = \frac{1}{3}$$

$$\omega = \frac{3}{10}$$

$$H\left(\frac{1}{3}\right) = 0.918$$

$$\text{info gain} = 1 - \left[\frac{7}{10} (0.985) + \frac{3}{10} (0.918) \right] = 0.0351$$

tail bushiness



$$P_1 = \frac{3}{4}$$

$$\omega = \frac{4}{10}$$

$$H\left(\frac{3}{4}\right) = 0.811$$

$$P_1 = \frac{2}{6}$$

$$\omega = \frac{6}{10}$$

$$H\left(\frac{2}{6}\right) = 0.918$$

$$\text{info gain} = 1 - \left[\frac{4}{10} (0.811) + \frac{2}{6} (0.918) \right] = 0.3696$$

Highest info gain from tail bushiness



$$H(p_{\text{root}}) = H(\frac{3}{4}) = 0.811$$

$$H(p_{\text{root}}) = H(\frac{2}{6}) = 0.918$$

	long	snout length	short	
wolf:	2		1	wolf: 2
not wolf:	1		0	not wolf: 0

$$P_i = \frac{2}{3}$$

$$w = \frac{3}{4}$$

$$H(\frac{2}{3}) = 0.918$$

$$P_i = \frac{1}{4}$$

$$w = \frac{1}{4}$$

$$H(1) = 0$$

$$\text{info gain} = 0.811 - \left[\frac{2}{3} (0.918) + 0 \right]$$

$$= 0.199$$

	long	snout length	short	
wolf:	2		0	wolf: 0
not wolf:	0		4	not wolf: 4

$$P_i = \frac{2}{2}$$

$$w = \frac{2}{6}$$

$$H(1) = 0$$

$$P_i = 0$$

$$w = \frac{4}{6}$$

$$H(0) = 0$$

$$\text{info gain} = 0.918 - [0+0]$$

$$= 0.918$$

	broad	face shape	narrow	
wolf:	2		1	wolf: 2
not wolf:	0		1	not wolf: 1

$$P_i = \frac{2}{2}$$

$$w = \frac{2}{4}$$

$$H(1) = 0$$

$$P_i = \frac{1}{2}$$

$$w = \frac{2}{4}$$

$$H(\frac{1}{2}) = 1$$

$$\text{info gain} = 0.811 - \left[0 + \frac{2}{4} (1) \right]$$

$$= 0.311$$

	broad	face shape	narrow	
wolf:	2		0	wolf: 0
not wolf:	3		1	not wolf: 1

$$P_i = \frac{2}{5}$$

$$w = \frac{5}{6}$$

$$H(\frac{2}{5}) = 0.971$$

$$P_i = 0$$

$$w = \frac{1}{6}$$

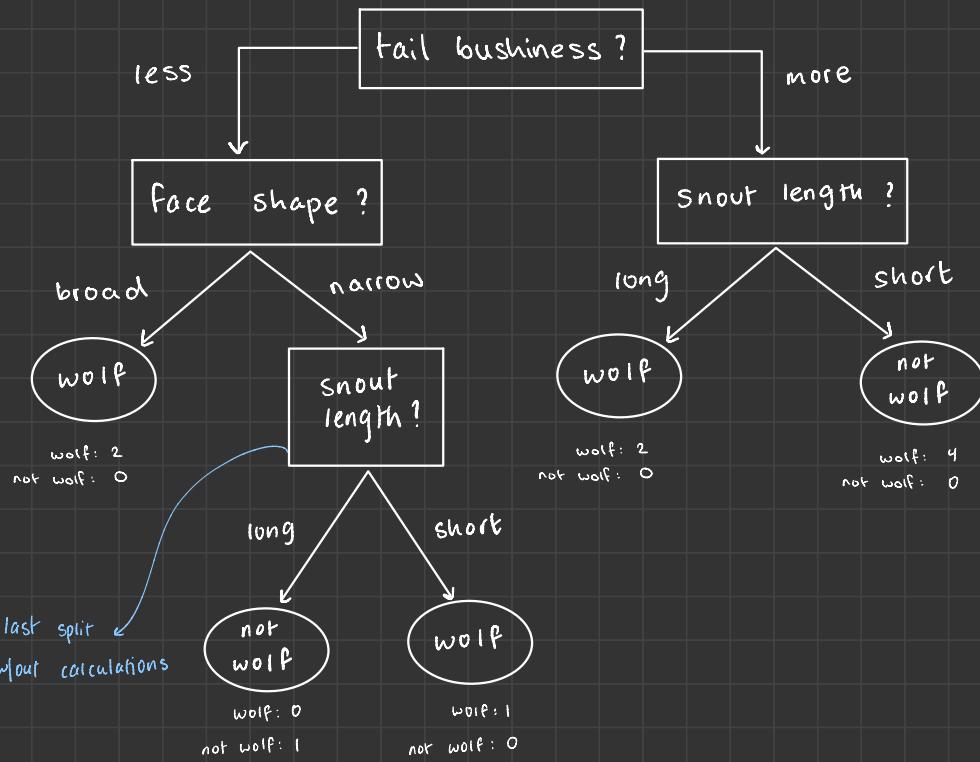
$$H(0) = 0$$

$$\text{info gain} = 0.918 - \left[\frac{5}{6} (0.971) + 0 \right]$$

$$= 0.109$$

higher info gain from
face shape

highest info gain from
snout length



— regression trees —

when target is numerical :

$$\text{variance} = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{instead of } H(p.)$$

10 samples

Snout Length	Face Shape	Tail Bushiness	Weight (lbs.)
Long	Broad	Less	15 ✓
Short	Narrow	Less	9.2 ✓
Short	Broad	More	7.2
Long	Narrow	Less	8.8
Long	Broad	Less	11
Long	Broad	More	18
Short	Narrow	More	8.4
Long	Broad	More	20
Short	Broad	More	7.6
Short	Broad	More	10.2

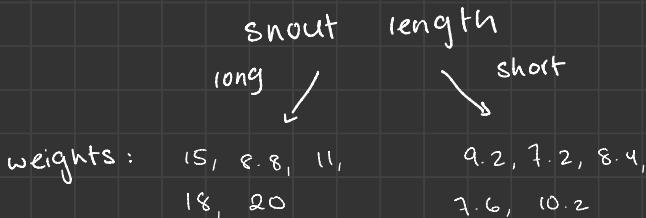
44.082

deciding root node only:

$$\bar{x} = \frac{115.4}{10} = 11.54$$

$$\sigma_{\text{root}}^2 = \frac{\sum (x_i - 11.54)^2}{10-1} = \frac{184.564}{9} = 20.51$$

→ here its
n-1



$$\sigma^2 : \frac{87.45}{5} = 17.49 \quad \frac{5.88}{5} = 1.18$$

→ here its n

$$\omega : \frac{5}{10} \quad \frac{5}{10}$$

$$\text{info gain} = 20.51 - \left[\frac{5}{10} (17.49) + \frac{5}{10} (1.18) \right] = 11.175$$

Other notes

interpreting ω_0 & ω_1 when the data was scaled:

$$y = \omega_0 + \omega_1 x \rightarrow \hat{y} = \hat{\omega}_0 + \hat{\omega}_1 \hat{x}$$

I Min Max Scaler

$$\text{scaled} \quad \text{original}$$
$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)} = \frac{x - \min(x)}{\text{range}(x)}$$

$$\Rightarrow \hat{y} = \hat{\omega}_0 + \hat{\omega}_1 \left[\frac{x - \min(x)}{\text{range}(x)} \right]$$

$$y = \underbrace{\omega_0}_{\hat{\omega}_0} + \underbrace{\frac{\hat{\omega}_1}{\text{range}(x)} \cdot x}_{\omega_1} - \underbrace{\frac{\hat{\omega}_1 \cdot \min(x)}{\text{range}(x)}}_{\omega_0}$$

$$\Rightarrow \omega_1 = \frac{\hat{\omega}_1}{\text{range}(x)}$$

$$\Rightarrow \omega_0 = \hat{\omega}_0 - \frac{\hat{\omega}_1 \cdot \min(x)}{\text{range}(x)}$$

2 Standard Scaler

$$y = \omega_0 + \omega_1 x \rightarrow \hat{y} = \hat{\omega}_0 + \hat{\omega}_1 \hat{x}$$

$$\begin{array}{c} \text{scaled} \\ \hat{x} = \frac{x - \text{mean}(x)}{\text{std}(x)} \end{array} \quad \begin{array}{c} \text{original} \\ = \frac{x - M}{\sigma} \end{array}$$

$$\Rightarrow \hat{y} = \hat{\omega}_0 + \hat{\omega}_1 \left[x \cdot \sigma + \mu \right]$$

$$y = \underbrace{\hat{\omega}_0}_{\omega_0} + \underbrace{(\hat{\omega}_1 \sigma)}_{\omega_1} x + \underbrace{\hat{\omega}_1 \mu}_{\omega_0}$$

$$\Rightarrow \hat{\omega}_1 = \frac{\hat{\omega}_1}{\sigma}$$

$$\Rightarrow \omega_0 = \hat{\omega}_0 + \hat{\omega}_1 \mu$$

x	y
2	10
4	15
6	20
8	25

OLS

(a) Scale

x	y
-1.34	-1.34
-0.45	-0.45
0.45	0.45
1.34	1.35

$$(b) \omega_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

