week 1 and week 2 → useless

week 3: Data quality exploration

types of features

1 Numerical (quantitative)
- continuos
- discrete

imputed using mean or median ← not good when data is skewed

- interval
  - no true zero value
- ratio
  - meaningful true zero

2 Categorical (qualitative)
- nominal
- ordinal

imputed using mode or median

Correlations

1 numerical VS numerical :

- Pearson → (-1, 1)
  - continuos variables
  - normal distributions
  - measures linearity
- Spearman → (-1, 1)
  - blw 2 continuos or 2 ordinal categorical
  - 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 independance

data preparation

1 categorical attributes

label / ordinal encoding

1st class	→	0
2nd class	→	1
⋮		
nth class	→	n-1

nominal / one-hot encoding

man	0	0	1
woman	0	1	0
child	1	0	0

2 numerical attributes

- min-max scaling

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

range: $[0, 1]$

- Standard scaler

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

μ: mean, σ: standard deviation

$$X \sim N(0, 1)$$

correlation examples

1 Pearson's correlation coefficient

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

cov: covariance, σ: standard deviation

X	Y
2	3
4	7
6	9
8	11

$\bar{X} = 5$

$\bar{Y} = 7.5$

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

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

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

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

$$\sigma_Y = \sqrt{\frac{\Sigma (y_i - \bar{y})^2}{N}} = \sqrt{\frac{35}{4}} = \frac{\sqrt{35}}{2}$$

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

$\Rightarrow$ strong positive linear relationship

2 Spearman's correlation coefficient

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

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

	X	Y	
1	97	195	1 → 1.5
2	99	195	2
3	101	210	5
4	103	205	3 → 3.5
5	105	205	4

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

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

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

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

means of class 1 & 0

$$\rho = \frac{M_1 - M_0}{\sigma} \times \sqrt{p_1 p_0}$$

σ: standard deviation of Y

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

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

$$p_1 = \frac{3}{6} = 0.5 \qquad p_0 = \frac{3}{6} = 0.5$$

$$\sigma = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} \qquad \bar{y} = \frac{248}{3} = 82.67$$

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

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

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

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

MSB: mean sum of square between groups

MSW: " " within groups

$$MSB = \frac{SSB}{df_B} = \frac{\sum n_i (\bar{x}_i - \bar{x})^2}{\#\text{ of groups} - 1}$$

SSB: sum of squares between; n_i: # of class i; $\bar{x}_i$: mean of class i; $\bar{x}$: grand mean; df_B: degrees of freedom

$$MSW = \frac{\sum SSW_i}{df_W} = \frac{\sum_j \sum_i (x_i - \bar{x}_j)^2}{\#\text{ of observations} - \#\text{ of groups}}$$

$j \in$ classes; $i \in$ each class; x_i: observation in class i; $\bar{x}_j$: class mean

Test Scores	Teaching Method	
85	A	] 3
90	A	
88	A	
75	B	] 2
78	B	
82	C	] 4
85	C	
87	C	
90	C	

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

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

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

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

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

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

$SSW_A = (85 - 87.69)^2 + \ldots = 12.67$

$SSW_B = 4.5$

$SSW_C = 34$

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

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

comparing F-statistic w/ critical value from table at p-value = 0.05 or 0.01 for df(x, y)
(col, row)

if (F-statistic $\geq$ critical) $\Rightarrow$ significant at $\alpha = 0.05 \mid 0.01$

df(2, 6)
$\alpha = 0.05 \rightarrow$ critical = 5.14 $\leq$ 9.796 $\Rightarrow$ 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 independance test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O: observed frequency; E: expected frequency

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

$\sum$: across the combination of two classes

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

r: # of classes in 1st group; c: # of classes in 2nd group

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

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

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 (χ^2 > critical) → significant

$\chi^2 = 27.977$
$\alpha = 0.05$
$df = 2$

critical = $5.991 \leq 27.977 \Rightarrow$ 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: $J(w) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2$

  $J(w)$: mean sum of square error; m: sample size; $h_w(x_i)$: prediction; y_i: true label

goal: $\min_w J(w)$

repeat until convergence: $w_j := w_j - \alpha \frac{d}{dw_j} J(w)$

α: learning rate; $\frac{d}{dw_j} J(w)$: 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:

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

$$\frac{\partial}{\partial w_0} J(w_0, w_1) = \frac{1}{\cancel{2}m} \sum_{i=1}^{m} \cancel{2} (\underbrace{w_0 + w_1 x_i}_{h_w(x_i)} - y_i) \cdot 1$$

$1 \rightarrow \frac{d}{dw_j} h_w(x_i)$

$$\frac{\partial}{\partial w_1} J(w_0, w_1) = \frac{1}{\cancel{2}m} \sum_{i=1}^{m} \cancel{2} (\underbrace{w_0 + w_1 x_i}_{h_w(x_i)} - y_i) \cdot x_i$$

$x_i \rightarrow \frac{d}{dw_j} h_w(x_i)$

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 $x \in \mathbb{R}^{m \times n}$

- $h_w(x) = w_0 + w_1 x_1 + \dots + w_n x_n$ (n = number of features)
- SGD: $w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_w(x) - y_i) \cdot x_{ij}$ (x_{ij}: sample i, feature j)
- convergence: when $\Delta J(w) \leq 10^{-3}$ after an iteration
- vector representation:

$w^T = [w_0 \quad w_1 \quad \dots \quad w_n]$ $w \in \mathbb{R}^{n+1}$ (n features + intercept)

$$x = \begin{bmatrix} x_0^{(1)} & \dots & x_0^{(m)} \\ x_1^{(1)} & \dots & \vdots \\ \vdots & & \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix}$$

all 1's for intercept

$x \in \mathbb{R}^{(n+1) \times m}$ (m samples)

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

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

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

or $J(w) = \frac{1}{2m} \left(w^T X - y \right)^T \left(w^T X - y \right)$

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

normal equation

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

Let $w \in \mathbb{R}^{(n+1) \times 1}$, $X \in \mathbb{R}^{m \times (n+1)}$ (changed from single sample example but whatever), $y \in \mathbb{R}^{m \times 1}$, then:

$$J(w) = \frac{1}{2m} \overbrace{(Xw - y)}^{m \times 1}{}^T (Xw - y)$$

$(n+1) \times m$ · $m \times (n+1)$ → $(n+1) \times (n+1)$

$$Xw \cdot Xw = w^T X^T X w$$

$1 \times (n+1)$ · $(n+1) \times 1$

scalar

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

$$Xw \cdot y = \underbrace{w^T X^T}_{1 \times m} y$$

$1 \times (n+1)$ · $(n+1) \times m$ · $m \times 1$

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

$$\frac{1}{m} \left(X^T X w - X^T y \right) = 0$$

$$X^T X w = X^T y$$

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

$n \times m$ · $m \times n$ → $n \times n$; $n \times m$ · $m \times 1$ → $n \times 1$

→ substitute & find optimal w vector

ordinary least squares

Special case of normal equation with one feature only :

$n = 1$

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

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

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

OLS example

x	y
1	2
2	3
3	5
4	7

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

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

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

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

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

$$\boxed{h_w(x) = 1.7x}$$

evaluation metrics

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}$$

(root mean square error)

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

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

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

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

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

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

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

→ asseses how well the model explains variance in the data

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

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(m-1)}{m-n-1}$$

→ decreases with the addition of irrelevant features

m: sample size; n: # of features

week 5: logistic regression

$$h_w(x) = \sigma(w^T x) \quad \text{s.t.} \quad 0 \le h_w(x) \le 1$$

(sigmoid; probability)

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

prediction threshold = 0.5

- class 1 if $h_w(x) \ge 0.5$
- 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)

$-\log(1) = 0$, $-\log(0) \to \infty$

$-\log(1-0) = 0$, $-\log(1-1) \to \infty$

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

- y: "activates" the corresponding cost when $y = 1$
- $(1-y)$: "activates" the corresponding cost when $y = 0$

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

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

goal: minimize $J(w)$

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

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

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

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

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

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

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

$$\frac{\partial}{\partial x} \sigma(f(x)) = \sigma(f(x))(1 - f(x)) \cdot f'(x)$$

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

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

x_j: feature j

↳ $x_0 = 1$ because intercept w_0

similarly:

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

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

$$\frac{\partial J(w)}{\partial w_j} = \frac{-1}{m} \sum_{i=1}^{m} \frac{y^{(i)}}{\sigma(w^T x)} \cdot \sigma(w^T x)\left(1 - \sigma(w^T x)\right) \cdot x_j^{(i)}$$

$$+ \frac{1 - y^{(i)}}{1 - \sigma(w^T x)} \cdot -\sigma(w^T x)\left(1 - \sigma(w^T x)\right) \cdot x_j^{(i)}$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{-1}{m} \sum_{i=1}^{m} y^{(i)}\left(1 - \sigma(w^T x)\right) x_j^{(i)} - \left(1 - y^{(i)}\right) \sigma(w^T x)\, x_j^{(i)}$$

$$y^{(i)} x_j^{(i)} - y^{(i)} x_j^{(i)} \sigma(w^T x) - x_j^{(i)} \sigma(w^T x) + y^{(i)} x_j^{(i)} \sigma(w^T x)$$

$$= \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} x_j^{(i)} - x_j^{(i)} \sigma(w^T x)$$

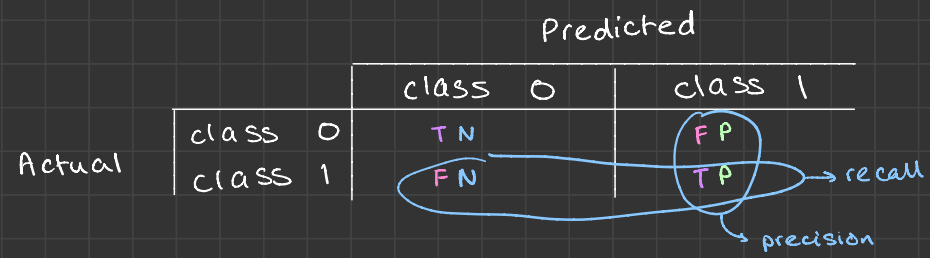
($\sigma(w^T x) = h_w(w^T x)$)

$$= \frac{-1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h_w(x^{(i)})\right) x_j^{(i)}$$

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

all of this is for binary classification, not multiclass

evaluation metrics



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

after targeting +ve, how many we correct / "precise" ?

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

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

!

similar to accuracy for +ve only

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

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.

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

- 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

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

$J(w)$: cost function

λ: regularization parameter

gradient descent:

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

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

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

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

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

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

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

normal equation

ridge (L2) regularization

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

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

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

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

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

$I_{4\times4}$

$m\times(n+1)$ $\quad 3\times3$ $\quad 3\times1$

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

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

lasso (L1) regression

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

$J(w)$: cost function

λ: regularization parameter

$|w_j|$: absolute value, not w_j^2

gradient descent:

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

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

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

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

pulls w_j closer to 0 during update → good for feature selection

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

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

logistic regression

ridge (L2) regularization

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

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

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

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

logistic regression

lasso (L1) regularization

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

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

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

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

week 6: decision trees

entropy — measure of impurity

$$H(p_1) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

n: # of classes; p_i: fraction of samples w class i

$H(p_1) = 0 \rightarrow$ 100% pure — 0% or 100% from 1 class

$H(p_1) = 1 \rightarrow$ 100% impure — 50% between 2 diff. classes

$$\text{entropy} = H(p_1) = -p_1 \log_2(p_1) - \underbrace{p_0}_{1-p_1} \log_2(\underbrace{p_0}_{1-p_0}) \rightarrow \text{binary classification}$$

$$\text{information gain} = H(p_1^{\text{root}}) - \left[w^{\text{left}} H(p_1^{\text{left}}) + w^{\text{right}} H(p_1^{\text{right}}) \right]$$

$H(p_1^{\text{root}})$: entropy at root; w^{left}: # of samples in left node

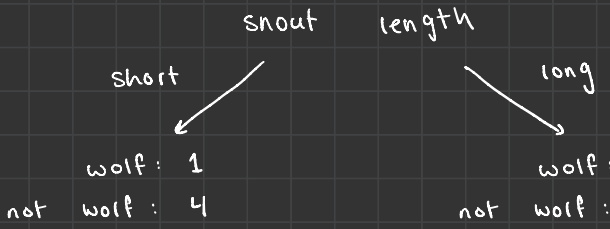
1. calculate information gain for all possible splits
2. choose highest information gain
3. split & repeat until:
   - $H(p_1) = 0$ for a node
   - max. depth reached
   - info. gain < threshold ∀ splits
   - # of samples in node < threshold

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

$$H(P_1^{root}) = -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right) = 1$$



$P_1 = 1/5$, $w = 5/10$

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

$P_1 = 4/5$, $w = 5/10$, $H(P_1) = 0.722$

$$\text{info. gain} = H(P_1^{root}) - \left[w^{left} H(P_1^{left}) + w^{right} H(P_1^{right})\right]$$

$$= 1 - \left[\frac{5}{10}(0.722) + \frac{5}{10}(0.722)\right]$$

$$= 0.278$$

face shape

broad → wolf: 4, not wolf: 3

narrow → wolf: 1, not wolf: 2

broad: $P_1 = 4/7$, $w = 7/10$, $H(4/7) = 0.985$

narrow: $P_1 = 1/3$, $w = 3/10$, $H(1/3) = 0.918$

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

tail bushiness

less → wolf: 3, not wolf: 1

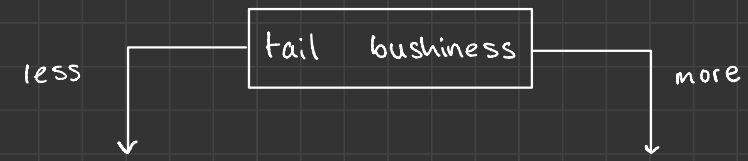
more → wolf: 2, not wolf: 4

less: $P_1 = 3/4$, $w = 4/10$, $H(3/4) = 0.811$

more: $P_1 = 2/6$, $w = 6/10$, $H(2/6) = 0.918$

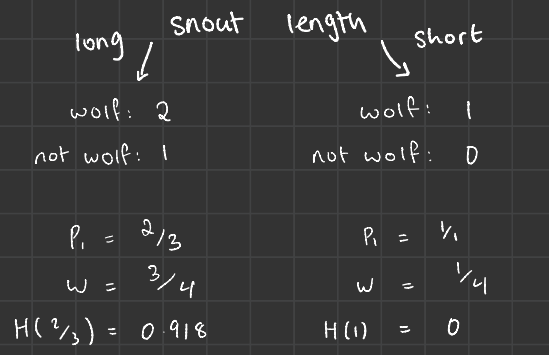
$$\text{info gain} = 1 - \left[\frac{4}{10}(0.811) + \frac{2}{6}(0.918) \right] = 0.3696$$

Highest info gain from tail bushiness



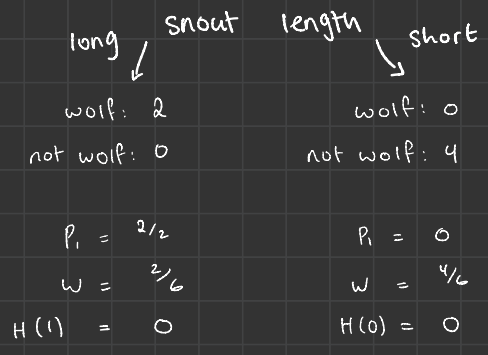
$H(P_1^{root}) = H(3/4) = 0.811$

$H(P_1^{root}) = H(2/6) = 0.918$



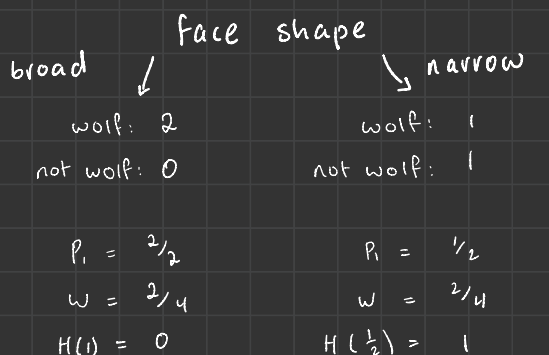
$\text{info gain} = 0.811 - \left[\frac{2}{3}(0.918) + 0\right]$

$= 0.199$



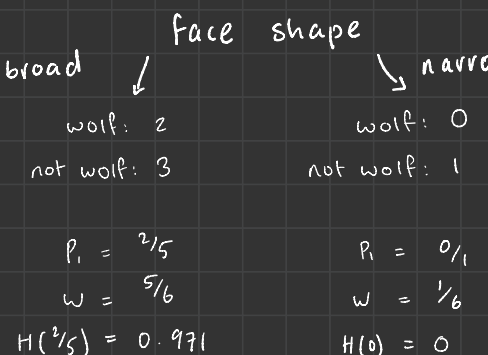
$\text{info gain} = 0.918 - [0 + 0]$

$= 0.918$



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

$= 0.311$

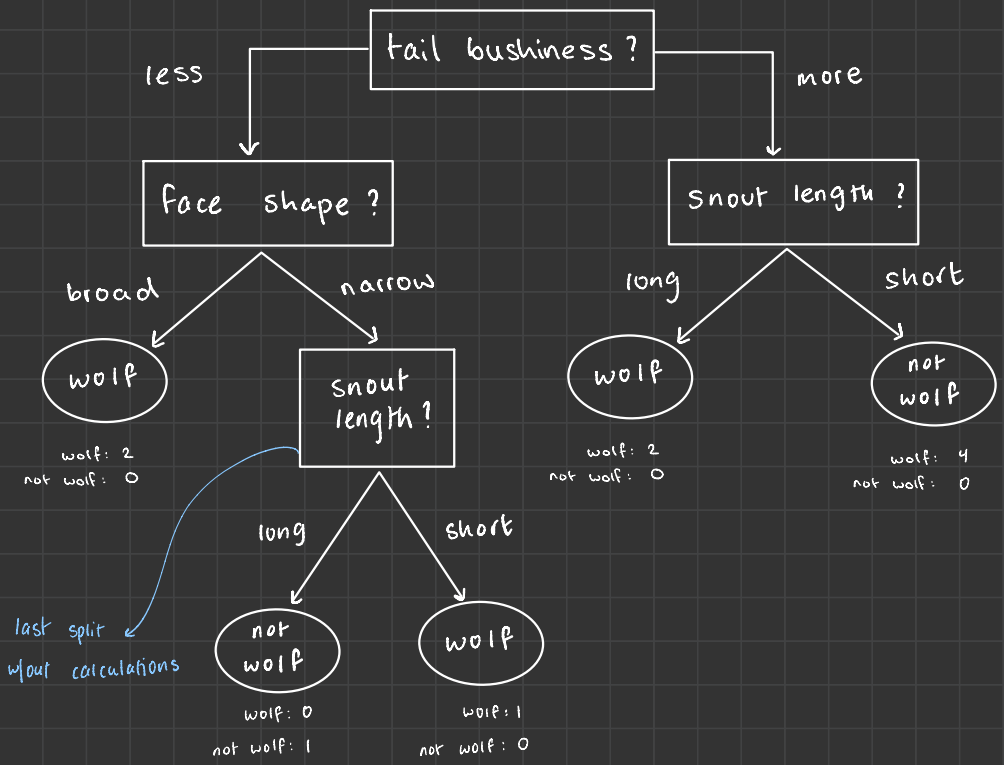


$\text{info gain} = 0.918 - \left[\frac{5}{6}(0.971) + 0\right]$

$= 0.109$

higher info. gain from face shape

highest info gain from snout length



regression trees

when target is numerical:

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

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:

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

$$\sigma^2_{root} = \frac{\Sigma (x_i - 11.54)^2}{10 - 1} = \frac{184.564}{9} = 20.51$$

→ here its n - 1

snout length

long ↙ short ↘

weights: 15, 8.8, 11, 18, 20 | 9.2, 7.2, 8.4, 7.6, 10.2

σ^2: $\frac{87.45}{5} = 17.49$ | $\frac{5.88}{5} = 1.18$

→ here its n

w: 5/10 | 5/10

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

single decision trees

- highly sensitive to small changes in data
- not very robust
- many decision trees would be more accurate
- single decision tree is more interpretable
- features are NOT scaled

ensemble trees

1 bagging:

- bootstrapping the data (sampling w/ replacement)
- train various trees w/ different data subsets for each
- aggregating results of various trees (voting)

- parallel training
- less computationally expensive
- less prone to overfitting

2 boosting:

- train one tree
- check predictions
- give more weight to misclassified samples
- train next tree accordingly

- sequential training of trees
- more computationally expensive
- more prone to overfitting

3 random forests

- at each node, take a random subset k from n features
- train on those features (i.e. decide on split)
- typically $k = \sqrt{n}$
- this could be done with bagging

4 XGBOOST

- subsequent trees trained on residuals of previous trees

other notes

interpreting w_0 & w_1 when the data was scaled:

$$y = w_0 + w_1 x \longrightarrow y = \hat{w}_0 + \hat{w}_1 \hat{x}$$

1 Min Max Scaler

scaled $\hat{x}$ = original x

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

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

$$y = \underbrace{\hat{w}_0}_{w_0} + \underbrace{\frac{\hat{w}_1}{\text{range}(x)}}_{w_1} \cdot x - \underbrace{\frac{\hat{w}_1 \cdot \min(x)}{\text{range}(x)}}_{w_0}$$

$$\Rightarrow w_1 = \frac{\hat{w}_1}{\text{range}(x)}$$

$$\Rightarrow w_0 = \hat{w}_0 - \frac{\hat{w}_1 \cdot \min(x)}{\text{range}(x)}$$

2 Standard Scaler

$$y = w_0 + w_1 x \longrightarrow y = \hat{w}_0 + \hat{w}_1 \hat{x}$$

scaled ↑ $\hat{x}$ = original ↑ $\frac{x - \text{mean}(x)}{\text{std}(x)} = \frac{x - \mu}{\sigma}$

$$\Rightarrow y = \hat{w}_0 + \hat{w}_1 \left[x \cdot \sigma + \mu \right] \quad (\hat{x})$$

$$y = \underbrace{\hat{w}_0}_{w_0} + \underbrace{(\hat{w}_1 \sigma)}_{w_1} x + \underbrace{\hat{w}_1 \mu}_{w_0}$$

$$\Rightarrow w_1 = \hat{w}_1 \sigma$$

$$\Rightarrow w_0 = \hat{w}_0 + \hat{w}_1 \mu$$