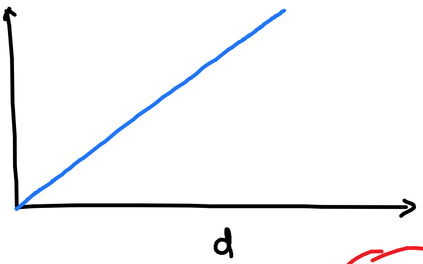
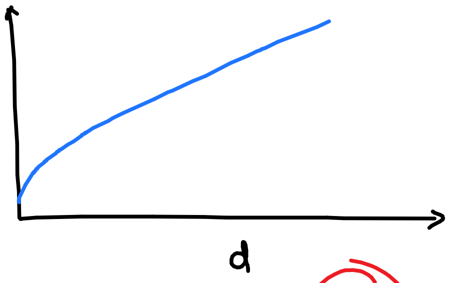
Homework 1

1) a) μ_{dist}



1/2

σ_{dist}



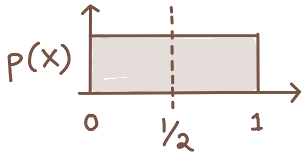
1/2

b) $z = (x-y)^2$

$x, y \sim$ independent r.v. from Unif $(0,1)$

Mean of $x, y = \mathbb{E}[x] = \frac{1}{2}$

Variance of $x = \mathbb{E}[(x-\mu)^2] = \frac{1}{12}$



Expected value of z

$$\mathbb{E}[z] = \mathbb{E}[(x-y)^2] = \mathbb{E}\left[\left\{\left(x-\tfrac{1}{2}\right) - \left(y-\tfrac{1}{2}\right)\right\}^2\right]$$

$$= \mathbb{E}\left[\left(x-\tfrac{1}{2}\right)^2 + \left(y-\tfrac{1}{2}\right)^2 + \left(x-\tfrac{1}{2}\right)\left(y-\tfrac{1}{2}\right)\right]$$

$$= \mathbb{E}\left[\left(x-\tfrac{1}{2}\right)^2\right] + \mathbb{E}\left[\left(y-\tfrac{1}{2}\right)^2\right] + \underbrace{\mathbb{E}\left[\left(x-\tfrac{1}{2}\right)\left(y-\tfrac{1}{2}\right)\right]}_{\text{x and y are independent of each other}}$$

same

$$= 2\,\mathbb{E}\left[\left(x-\underset{\tilde{\mu}_x}{\tfrac{1}{2}}\right)^2\right] + \mathbb{E}\left[\left(x-\tfrac{1}{2}\right)\right]\mathbb{E}\left[\left(y-\tfrac{1}{2}\right)\right]$$

Note: $\mathbb{E}\left[\left(x-\frac{1}{2}\right)\right]$

$= \mathbb{E}[x] - \frac{1}{2}$

$= \frac{1}{2} - \frac{1}{2} = 0$

$$= 2 \times \frac{1}{12} = \frac{1}{6}$$

$$\boxed{\mathbb{E}[z] = 1/6}$$

0.75

$$\mathbb{E}[(z-\mu_z)^2] = \mathbb{E}[z^2] - \mu_z^2$$

$$= \mathbb{E}[(x-y)^4] - (\mathbb{E}[z])^2$$

$$= \mathbb{E}[x^4 + y^4 - 4x^3y - 4xy^3 + 6x^2y^2] - \frac{1}{36}$$

$$= \mathbb{E}[x^4] + \mathbb{E}[y^4] - 4\,\mathbb{E}[x^3y] - 4\,\mathbb{E}[xy^3] + 6\,\mathbb{E}[x^2y^2] - \frac{1}{36}$$

$$\mathbb{E}[X^4] = \int_0^1 x^4 \frac{1}{(1-0)}\, dx = \left.\frac{x^5}{5}\right|_0^1 = 1/5$$

$$\mathbb{E}[X^3 Y] = \mathbb{E}[X^3]\,\mathbb{E}[Y] \quad \text{(due to independence)}$$

$$= \left.\frac{x^4}{4}\right|_0^1 \times \left.\frac{y^2}{2}\right|_0^1 = 1/8$$

$$\mathbb{E}[X Y^3] = \left.\frac{x^2}{2}\right|_0^1 \times \left.\frac{y^4}{4}\right|_0^1 = 1/8$$

$$\mathbb{E}[Y^4] = \left.\frac{y^5}{5}\right|_0^1 = 1/5 \qquad \mathbb{E}[X^2 Y^2] = \left.\frac{x^3}{3}\right|_0^1 \times \left.\frac{y^3}{3}\right|_0^1 = \frac{1}{9}$$

$$\mathbb{E}[(Z-\mu_Z)^2] = \mathbb{E}[X^4] + \mathbb{E}[Y^4] - 4\,\mathbb{E}[X^3 Y] - 4\,\mathbb{E}[XY^3] + 6\,\mathbb{E}[X^2Y^2] - \frac{1}{36}$$

$$= \frac{1}{5} + \frac{1}{5} - 4\left(\frac{1}{8}\right) - 4\left(\frac{1}{8}\right) + 6\left(\frac{1}{9}\right) - \frac{1}{36}$$

$$= \frac{2}{5} - 1 + \frac{2}{3} - \frac{1}{36} = \frac{7}{180}$$

$$\boxed{\mathrm{Var}(Z) = \mathbb{E}[(Z-\mu_Z)^2] = \frac{7}{180}}$$

1.25

c) $$\mathbb{E}[S] = \mathbb{E}[Z_1 + Z_2 + \cdots + Z_d]$$

$$= \mathbb{E}\left[\sum_{i=1}^{d} Z_i\right] = \mathbb{E}\left[\sum_{i=1}^{d} (X_i - Y_i)^2\right]$$

$$= \sum_{i=1}^{d} \mathbb{E}[(X_i - Y_i)^2]$$

We know that $\mathbb{E}[(X_i - Y_i)^2] = 1/6$ and that $\mathbb{E}[(X_i - Y_i)^2] = \mathbb{E}[(X_j - Y_j)^2]$ $\forall\, i,j \in d$

$$\therefore\ \mathbb{E}[S] = \sum_{i=1}^{d} \mathbb{E}[(X_i - Y_i)^2] = \sum_{i=1}^{d} (1/6) = d/6$$

$$\boxed{\mathbb{E}[S] = d\,\mathbb{E}[Z] = d/6}$$

0.5

Similarly, we calculate variance of S

$$\mathrm{Var}[S] = \mathrm{Var}[Z_1 + Z_2 + \cdots + Z_d]$$

$$= \sum_{i=1}^{d} \mathrm{Var}(Z_i) \quad \text{[since } Z_i\text{'s are independent]}$$

$$= d\,\mathrm{Var}(Z) = 7d/180$$

0.5

(Note $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Cov}(X,Y)$

If X and Y are independent $\Rightarrow \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$)

You can further use Markov's inequality to prove that points in higher dimensions are far apart

Markov's inequality says

$$p\left(|z - \mathbb{E}[z]| \geq a\right) \leq \frac{\mathrm{Var}[z]}{a^2}$$

or,

$$p\left(|s - \mathbb{E}[s]| \geq a\right) \leq \frac{\mathrm{Var}[s]}{a^2}$$

$$\Rightarrow \underbrace{p\left(\left|s - d/6\right| \geq a\right)}_{} \leq \frac{7d}{180a^2}$$

Note $s = \|\underline{x} - \underline{y}\|_2^2$ represents the distance between two points lying in d-dimensional space

probability that this distance $\left|s - \frac{d}{6}\right|$ is greater than 'a'

$\left|s - d/6\right| \rightarrow$ is also a distance

Say $a = 1$

For $d = 1$

$$p\left(\left|s - 1/6\right| \geq 1\right) \leq \frac{7}{180}$$

In 1-D, the chances of the distance between 2 points exceeding a certain values is less

For $d = 5$

$$p\left(\left|s - 5/6\right| \geq 1\right) \leq \frac{35}{180}$$

For $d = 10$

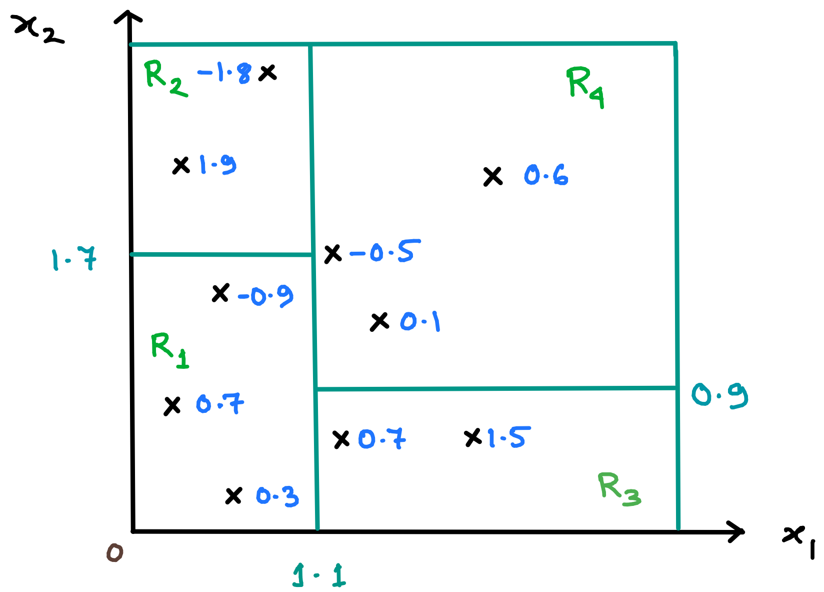
$$p\left(\left|s - 10/6\right| \geq 1\right) \leq \frac{70}{180}$$

In 10-D, the chances of the distance between 2 points exceeding a certain values is much more

Hence, we find that with increasing dimension, the distance between points increases, and most points in higher dimensions are quite far apart!

3) Regression Tree

a)



1

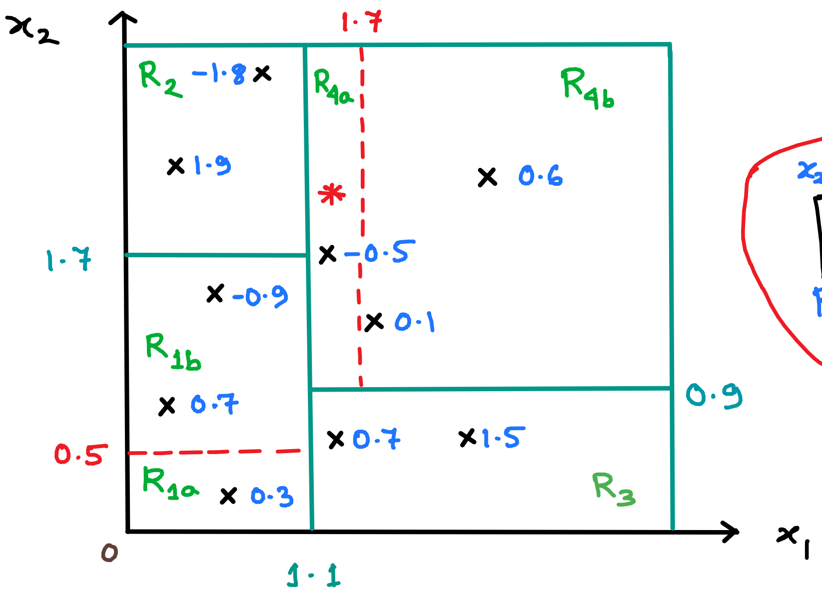
b) Since $x_{1,*} = 1.5 > 1.1$ and $x_{2,*} = 1.8 > 0.9$, the test points belongs to region R_4.

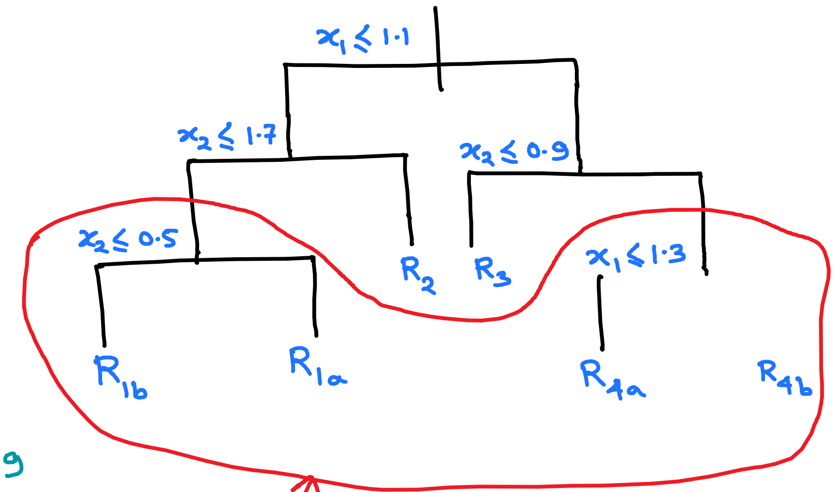
The mean of the training point output in R_4 is $\hat{y}_{R_4} = 0.0667$

Therefore, the prediction becomes $\hat{y}_* = 0.067$ 0.5

c) There could be many possibilities of creating a deeper tree.

One example could be





change must happen at this level

2

d) Based on the above tree, x_* belongs to region R_{4a},

0.5

thus $\hat{y}_* = \hat{y}_{R_{4a}} = -0.5$

5)

a) $\underline{y} = \underline{\underline{X}}\,\underline{\theta} + \underline{\epsilon}, \quad \underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_N)$

I_N is an identity matrix of size $N \times N$

N – size of training data

- The likelihood turns out to be Gaussian

$$P(\underline{y} \mid \underline{\underline{X}}\,\underline{\theta}) = \mathcal{N}(\underline{\underline{X}}\,\underline{\theta}, \sigma^2 I_N)$$

$$= \frac{1}{(2\pi)^{N/2}\,|\sigma^2 I_N|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\underline{y} - \underline{\underline{X}}\,\underline{\theta})^T(\underline{y} - \underline{\underline{X}}\,\underline{\theta})\right)$$

0.5

- Log-likelihood

$$\ln p(\underline{y} \mid \underline{\underline{X}}\,\underline{\theta}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{(\underline{y} - \underline{\underline{X}}\,\underline{\theta})^T(\underline{y} - \underline{\underline{X}}\,\underline{\theta})}_{\text{dependence on } \underline{\theta}}$$

To maximize the log-likelihood, we take derivative w.r.t. $\underline{\theta}$ and set it to zero

$$\frac{\partial}{\partial\underline{\theta}} \ln p(\underline{y} \mid \underline{\underline{X}}\,\underline{\theta}) = \frac{1}{2\sigma^2}\, 2\, \underline{\underline{X}}^T(\underline{y} - \underline{\underline{X}}\,\underline{\theta}) = 0$$

$$\underline{\underline{X}}^T(\underline{y} - \underline{\underline{X}}\,\underline{\theta}) = 0$$

$$\Rightarrow \underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\theta} = \underline{\underline{X}}^T\underline{y}$$

If $\underline{\underline{X}}^T\underline{\underline{X}}$ is invertible, then

$$\hat{\underline{\theta}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{y}$$

0.5

b) In practice, $\underline{\underline{X}}$ is a tall matrix with more rows than columns

The columns of matrix $\underline{\underline{X}}$ denote the different input features

If $\underline{\underline{X}}^T\underline{\underline{X}}$ is not invertible $\rightarrow$ $\underline{\underline{X}}$ is rank-deficient

0.5 $\rightarrow$ In practice, it means some input features are redundant

7) Logistic function, $h(x) = \dfrac{e^x}{1+e^x}$

0.5

a) $$\frac{dh(x)}{dx} = \frac{e^x(1+e^x) - e^x \cdot e^x}{(1+e^x)^2} = \frac{e^x(1+e^x-e^x)}{(1+e^x)^2}$$

$$= \frac{e^x}{1+e^x} \cdot \frac{1}{1+e^x}$$

$$= \left(\frac{e^x}{1+e^x}\right) \cdot \left(1 - \frac{e^x}{1+e^x}\right)$$

$$= h(x) \cdot (1 - h(x))$$

b) We will now consider the two classes as $\{0,1\}$ (instead of $\{-1,1\}$)

Treat

$$p(y=1 \mid \underline{x}; \underline{\theta}) = h(\underline{x}^T\underline{\theta}) \; ; \quad p(y=0 \mid \underline{x}; \underline{\theta}) = 1 - h(\underline{x}^T\underline{\theta})$$

$$= \frac{e^{\underline{x}^T\underline{\theta}}}{1+e^{\underline{x}^T\underline{\theta}}} \qquad\qquad = \frac{1}{1+e^{\underline{x}^T\underline{\theta}}}$$

Log-Likelihood for a data pair $\{\underline{x}_i, y_i\}$

$$\ln p(y_i \mid \underline{x}_i ; \underline{\theta}) = \begin{cases} \ln h(\underline{x}_i^T\underline{\theta}) & \text{if } y_i = 1 \\ \ln(1 - h(\underline{x}_i^T\underline{\theta})) & \text{if } y_i = 0 \end{cases}$$

To make the expression more compact, we write

$$\ln p(y_i \mid \underline{x}_i ; \underline{\theta}) = y_i \ln h(\underline{x}_i^T\underline{\theta}) + (1-y_i) \ln(1 - h(\underline{x}_i^T\underline{\theta}))$$

The log-likelihood for entire training data is

1

$$\ln p(y_1, \ldots, y_N \mid \underline{x}_1, \ldots, \underline{x}_N) = \sum_{i=1}^{N} y_i \ln h(\underline{x}_i^T\underline{\theta}) + (1-y_i) \ln(1 - h(\underline{x}_i^T\underline{\theta}))$$

c)

$$\ln p(y_1, \dots, y_N \mid \underline{x}_1, \dots, \underline{x}_N) = \sum_{i=1}^{N} y_i \ln h(\underline{x}_i^T \underline{\theta}) + (1-y_i) \ln(1-h(\underline{x}_i^T \underline{\theta}))$$

$$\frac{d}{d\underline{\theta}} \left[y_i \ln \underbrace{h(\underline{x}_i^T \underline{\theta})}_{h} + (1-y_i) \ln(1-\underbrace{h(\underline{x}_i^T \underline{\theta})}_{h}) \right]$$

$$= y_i \frac{1}{h} \left(\frac{dh}{d\underline{\theta}} \right) \underline{x}_i + (1-y_i) \frac{1}{1-h} \left(-\frac{dh}{d\underline{\theta}} \right) \underline{x}_i$$

Using the relation $\frac{dh}{d\underline{\theta}} = h(1-h)$

$$= y_i (1-h) \underline{x}_i - (1-y_i) h \underline{x}_i$$

$$= y_i \underline{x}_i - \cancel{y_i h \underline{x}_i} - h \underline{x}_i + \cancel{y_i h \underline{x}_i}$$

$$= (y_i - h) \underline{x}_i$$

$$= \underbrace{(y_i - h(\underline{x}_i^T \underline{\theta})) \underline{x}^{(i)}}$$

Therefore,

$$\frac{dL}{d\underline{\theta}} = \frac{d}{d\underline{\theta}} \ln p(\underline{y} \mid \underline{X}; \underline{\theta}) = \sum_{i=1}^{N} (y_i - h(\underline{x}_i^T \underline{\theta})) \underline{x}_i^T \qquad 1$$

d) Differentiating further,

$$\underbrace{\frac{d^2 \ln p(y_i \mid \underline{x}_i; \underline{\theta})}{d\underline{\theta}\, d\underline{\theta}^T}}_{\substack{p \times p \\ \text{matrix}}} = \frac{d}{d\underline{\theta}^T} (y_i - h(\underline{x}_i^T \underline{\theta})) \underline{x}_i$$

$$= -\frac{dh}{d\underline{\theta}^T} \underline{x}_i \underline{x}_i^T$$

$\underline{\theta} \in \mathbb{R}^p$

$$= -h(1-h) \underbrace{\underline{x}_i \underline{x}_i^T}_{p \times p} \qquad 1$$

$$\frac{d^2 L}{d\underline{\theta}\, d\underline{\theta}^T} = -\sum_{i=1}^{N} h(\underline{x}_i^T \underline{\theta}) (1-h(\underline{x}_i^T \underline{\theta})) \underline{x}_i \underline{x}_i^T$$