FoDA, Semester

L29: Review

1. Consider the random variables X and Y described by the joint probability table.

	$X = 1$	$X = 2$	$X = 3$	
$Y = 1$	0.10	0.05	0.10	
$Y = 2$	0.30	0.25	0.20	$= 0.75$
	0.40	0.3	0.3	

Derive the following values

(a) $\mathbf{Pr}(X = 1)$ $= 0.4$

(b) $\mathbf{Pr}(X = 2 \cap Y = 1)$ $= 0.05$

(c) $\mathbf{Pr}(X = 3 \mid Y = 2)$ $= \frac{0.2}{0.75}$

Compute the following probability distributions.

$X=1$	$X=2$	$X=3$	
0.4	.3	.3	= sum 1

(d) What is the marginal distribution for X?

(e) What is the conditional probability for Y, given that $X = 2$?

$Y=1$	$0.05 / 0.3$
$Y=2$	$0.25 / 0.3$

Answer the following question about the joint distribution.

(f) Are random variables X and Y independent? $\rightarrow$ NO

(g) Is $\mathbf{Pr}(X = 1)$ independent of $\mathbf{Pr}(Y = 1)$?

$\hookrightarrow$ Yes

$Pr[X=1] = 0.4$

$Pr[Y=1] = 0.25$

$Pr[X=1 \cap Y=1] = 0.1$

R.V. X, Y independent if all events $A \in X$, $B \in Y$ are independent

$$\underbrace{Pr[X=2 \cap Y=2]}_{0.25} \neq \underbrace{Pr[X=2]}_{0.3} \cdot \underbrace{P[Y=2]}_{0.75}$$

2. Consider two models M_1 and M_2, where from prior knowledge we believe that $\mathbf{Pr}(M_1) = 0.25$ and $\mathbf{Pr}(M_2) = 0.75$. We then observe a data set D. Given each model we assess the likelihood of seeing that data given the model as $\mathbf{Pr}(D \mid M_1) = 0.5$ and $\mathbf{Pr}(D \mid M_2) = 0.01$. Now that we have the data, which model is has a higher probability of being correct?

$$\boxed{\Pr(M_1 \mid D)} = \frac{\Pr[D \mid M_1] \cdot \Pr(M_1)}{\cancel{\Pr(D)}}$$

large

$$\propto \Pr[D \mid M_1] \cdot \Pr(M_1)$$

$$\propto (0.5) \cdot (0.25)$$

$$\propto 0.125$$

$$\Pr(M_2 \mid D) \propto \Pr[D \mid M_2] \cdot P(M_2)$$

$$\propto (0.01)(0.75)$$

$$\propto \boxed{(0.0075)}$$

3. Assume I observe 3 data points x_1, x_2, and x_3 drawn iid from an unknown distribution. Given a model M, I can calculate the likelihood this each data point as $\textbf{Pr}(x_1 \mid M) = 0.5$, $\textbf{Pr}(x_2 \mid M) = 0.1$, and $\textbf{Pr}(x_3 \mid M) = 0.2$. What is the likelihood of seeing all of these data points, given the model M: $\textbf{Pr}(x_1, x_2, x_3 \mid M)$?

$$Pr(x_1, x_2, x_3 \mid M) = Pr(x_1 \mid M) \cdot Pr(x_2 \mid M) \cdot Pr(x_3 \mid M)$$

$$= (0.5)(0.1)(0.2)$$

$$= 0.01$$

4. Consider a pdf f so that a random variable $X \sim f$ has expected value $\mathbf{E}[X] = 3$ and variance $\mathbf{Var}[X] = 10$. Now consider $n = 10$ iid random variables $X_1, X_2, \ldots, X_{10}$ drawn from f. Let $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

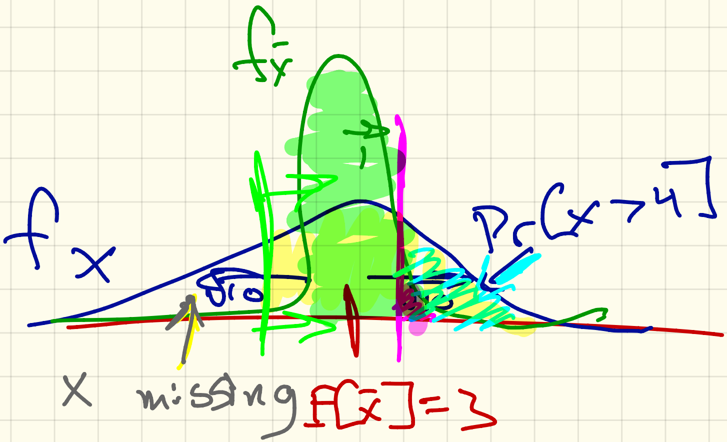
(a) What is $\mathbf{E}[\bar{X}]$? $= 3$

(b) What is $\mathbf{Var}[\bar{X}]$? $\mathrm{Var}[X]/10 = 1$

(c) What is the standard deviation of $\bar{X}$? $\sqrt{\mathrm{Var}[\bar{x}]} = \sqrt{1} = 1$

(d) Which is larger $\mathbf{Pr}[X > 4]$ or $\mathbf{Pr}[\bar{X} > 4]$?

(e) Which is larger $\mathbf{Pr}[X > 2]$ or $\mathbf{Pr}[\bar{X} > 2]$?

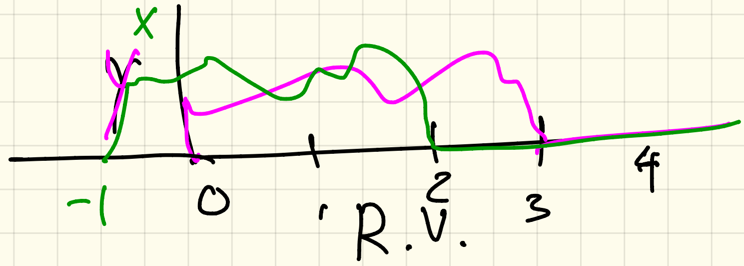


$Pr[x > 4] > Pr[\bar{x} > 4]$

5. Let X be a random variable that you know is in the range $[-1, 2]$ and you know has expected value of $\mathbf{E}[X] = 0$. Use the Markov Inequality to upper bound $\mathbf{Pr}[X > 1.5]$?
(Hint: you will need to use a change of variables.)

R.V. > 0

$Y = X + 1$



$E[Y] = E[X+1]$
$= E[X] + 1$

$\Pr[X^{+1} > 1.5^{+1}] = \Pr[Y > 2.5]$

M.I. $\leq \frac{E[Y]}{2.5} = \frac{E[X]+1}{2.5} = \frac{0+1}{2.5} = \frac{1}{2.5}$

$= 0.4$

$\Pr[X > 1.5] \leq 0.4$

6. Consider a matrix

$$A = \begin{bmatrix} 2 & 2 & 3 \\ -2 & 7 & 4 \\ -3 & -3 & -4.5 \\ -8 & 2 & 3 \end{bmatrix}. \begin{bmatrix} 2 \\ -2 \\ -3 \\ 1817 \end{bmatrix}$$

(a) Add a column to A so that it is invertable.

(b) Remove a row from A so that it is invertable.

(c) Is AA^T invertable? A is 4×3 so $AA^T = 4 \times 4$ $\quad \operatorname{rank}(AA^T) \le \min(\operatorname{rank}(A), \operatorname{rank}(A^T)) \le 3$

(d) Is $A^T A$ invertable? $\rightarrow$ No

$A^T A$ is 3×3

$\hookrightarrow$ Yes.

7. Consider two vectors $u = (0.5, 0.4, 0.4, 0.5, 0.1, 0.4, 0.1)$ and $v = (-1, -2, 1, -2, 3, 1, -5)$.

0.25 0.16 0.01

(a) Check if u or v is a unit vector.

(b) Calculate the dot product $\langle u, v\rangle$. $= 0$

(c) Are u and v orthogonal? $\rightarrow$ Yes

$$\begin{array}{ccccccc} 0.5 & .4 & .4 & .5 & .1 & .4 & .1 \\ \hline (-0.5) & (-.8) & (.4) & (-1) & (.3) & (.4) & (-.5) \end{array}$$

add up $\sum_j$ $= 0$

unit vector r has $\|r\|_2 = 1$

$$\sqrt{\sum_{j=1}^{d} r_j^2} = 1$$

$$2(0.25) + 3(0.16) + 2(0.01)$$

$$0.5 + 0.48 + 0.02 = 1$$

$$\sqrt{1} = 1$$

8. Consider a matrix $A \in \mathbb{R}^{n\times 4}$. Each row represents a customer (there are n customers in the database). The first column is the age of the customer in years, the second column is the number of days since the customer entered the database, the third column is the total cost of all purchases ever by the customer in dollars, and the last column is the total profit in dollars generated by the customer.

For each of the following operations, decide if it is **reasonable** or **unreasonable**.

(a) Run simple linear regression using the first three columns to build a model to predict the fourth column.

(b) Use k-means clustering to group the customers into 4 types using Euclidean distance between rows as the distance. unreasonable

(c) Use PCA to find the best 2-dimensional subspace, so we can draw the customers in a $\mathbb{R}^2$ in way that has the $\boxed{\text{least projection error.}}$ uses $\|a_i - \pi_F(a_i)\| \rightarrow$ NO

(d) Use the linear classification to build a model based on the first three columns to predict if the customer will make a profit $+1$ or not -1.

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

$$d(a_i, a_j) = \|a_i - a_j\|$$

$\hookrightarrow$ does not make sense

9. Consider a data set (X, y) where $X \in \mathbb{R}^{n\times 3}$ we decompose into a test and a training data set $(X_{\text{train}}, y_{\text{train}})$. Assume that X_{train} is not just a subset of X, but also pads prepends a columns of all 1s. We build a linear model

$$\alpha = (X_{\text{train}}^T X_{\text{train}})^{-1} X_{\text{train}}^T y_{\text{train}}.$$

where $\alpha \in \mathbb{R}^{\overset{4}{3}}$. The remaining two testing data points are $\boxed{(x_1, y_1)}$ and (x_2, y_2), where $x_1, x_2 \in \mathbb{R}^3$. Explain (write a mathematical expression) to use this test data to estimate the generalization error. That is, if one new data point arrives x, how much squared error would we expect the model α to have compared to the unknown true value y?

$$\left(\left(\langle \alpha, (1; x_1)\rangle - y_1\right)^2 + \left(\langle \alpha, (1; x_2)\rangle - y_2\right)^2 \right) \cdot \frac{1}{2}$$

$$\boxed{\frac{1}{k} \sum_{j=1}^{k} \left(\langle \alpha, (1; x_j)\rangle - y_j\right)^2}$$

10. Consider a function $f(x, y)$ with gradient $\nabla f(x, y) = (x - 1, 2y + x)$. Starting at a value $(x = 1, y = 2)$, and a learning rate of $\gamma = 1$, execute one step of gradient descent.

$$\nabla f(x=1, y=2) = (1-1, 2(2)+1)$$
$$= (0, 5)$$

$$\alpha = (x=1, y=2)$$

$$\alpha_f = \alpha - \gamma \nabla f(\alpha)$$
$$= (1, 2) - 1 \cdot (0, 5)$$
$$= (1, -3)$$

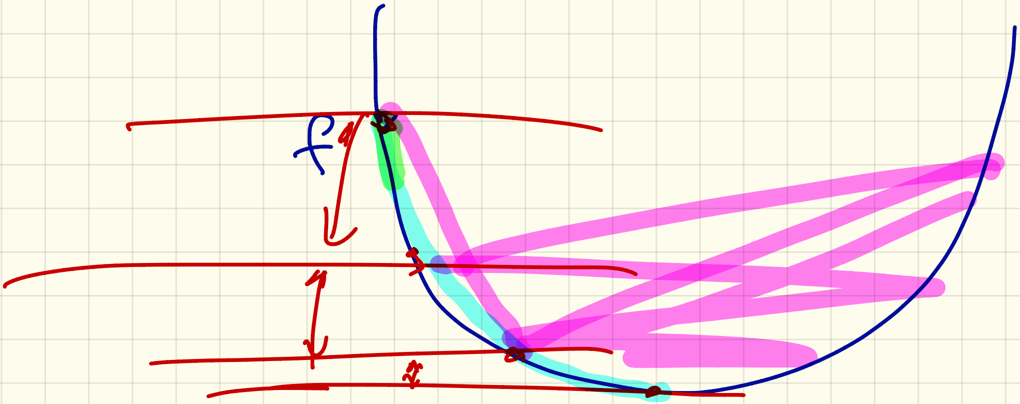
11. Consider running gradient descent with a fixed learning rate γ. For each of the following, we plot the function value over 10 steps (the function is different each time). Decide whether the learning rate is probably **too high**, **too low**, or **about right**.

(a) f_1: 100, 99, 98, 97, 96, 95, 94, 93, 92, 91 too low

(b) f_2: 100, 50, 75, 60, 65, 45, 75, 110, 90, 85 too high

(c) f_3: 100, 80, 65, 50, 40, 35, 31, 29, 28, 27.5, 27.3 right

(d) f_4: 100, 80, 60, 40, 20, 0, -20, -40, -60, -80, -100 too low



12. Consider a matrix $A \in \mathbb{R}^{8\times 4}$ with squared singular values $\sigma_1^2 = 10$, $\sigma_2^2 = 5$, $\sigma_3^2 = 2$, and $\sigma_4^2 = 1$.

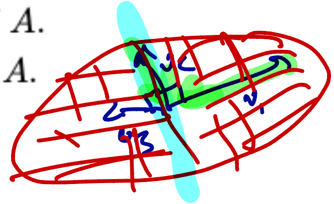
$\|A\|_F^2 = \sum_{j=1}^{4} \sigma_j^2 = \boxed{\cancel{10} + \cancel{5}} + 2 + 1 = 18$

(a) What is the rank of A? 4

(b) What is $\|A - A_2\|_F^2$, where A_2 is the best rank-2 approximation of A. $=3$

(c) What is $\|A - A_2\|_2^2$, where A_2 is the best rank-2 approximation of A. $=2$

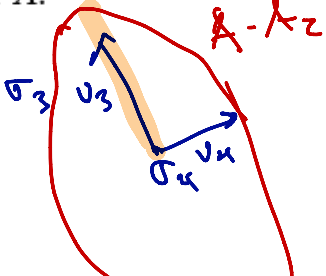
$\|A\|_2^2 = \sigma_1^2 = 10$

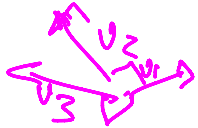


(d) What is $\|A\|_2^2$? $=10$

(e) What is $\|A\|_F^2$? $=18$

Let v_1, v_2, v_3, v_4 be the right singular vectors of A.





(f) What is $\|Av_2\|^2$? $=5$

(g) What is $\langle v_1, v_3\rangle$? $=0$

(h) What is $\|v_4\|$? $=1$

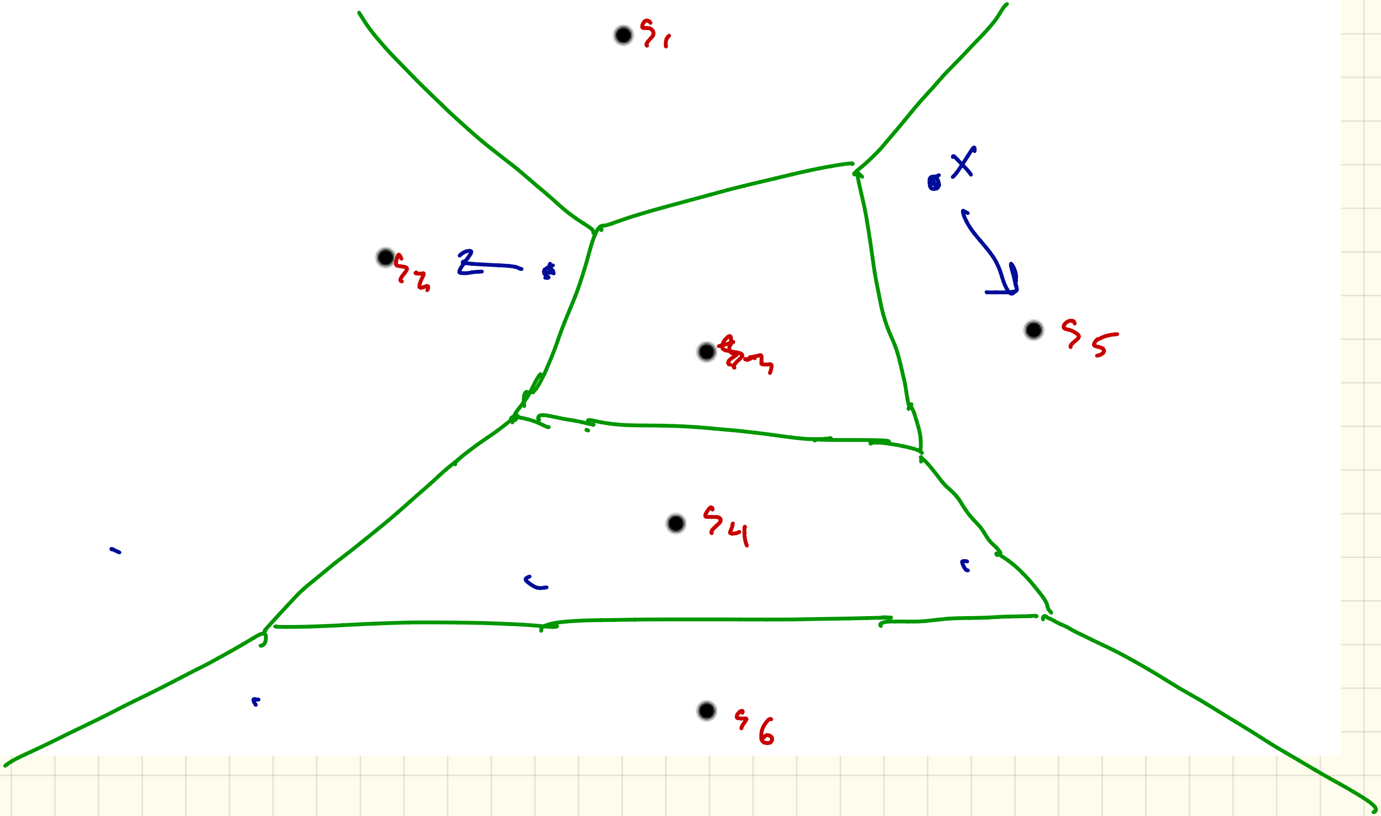
Let $a_1 \in \mathbb{R}^4$ be the first row of A.

(i) Write a_1 in the basis defined by the right singular vectors of A.

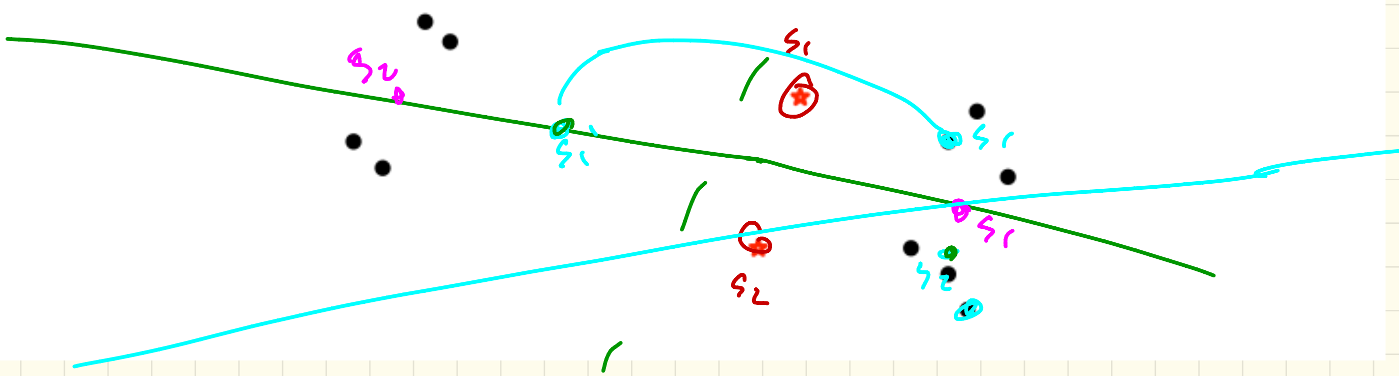
$V^T a_1 = (\langle v_1, a_1\rangle, \langle v_2, a_1\rangle, \langle v_3, a_1\rangle, \langle v_4, a_1\rangle)$

$A = \sum_{j=1}^{d=4} \sigma_j u_j v_j^T$ $\qquad A_2 = \sum_{j=1}^{2} \sigma_j u_j v_j^T$ $\qquad A - A_2 = \sum_{j=3}^{4} \sigma_j u_j v_j^T$

13. Draw the Voronoi diagram of the following set of points.



14. What should you do, if running Lloyd's algorithm for k-means clustering ($k = 2$), and you reach this scenario, where the algorithm terminates? (The black circles are data points and red stars are the centers).

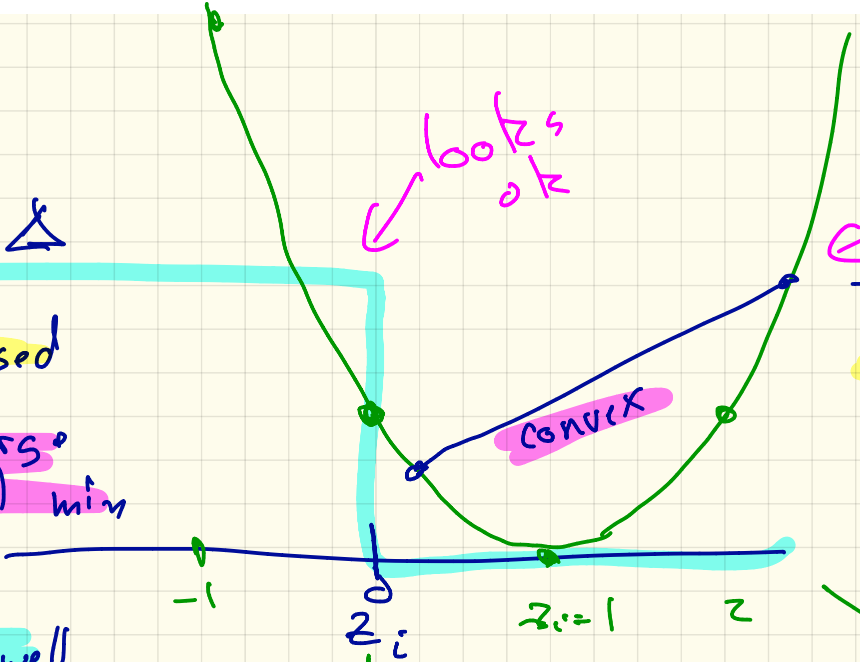


→ Random restart

a) $s_j \leftarrow \text{average cluster}(X_j)$

b) cluster X_j = all points x s.t. $\phi_S(x) = s_j$

15. Consider the following "loss" function. $\ell_i(z_i) = (1 - z_i)^2/2$, where for a data point (x_i, y_i) and prediction function f, then $z_i = y_i \cdot f(x_i)$. Predict how this might work within a gradient descent algorithm for classification.



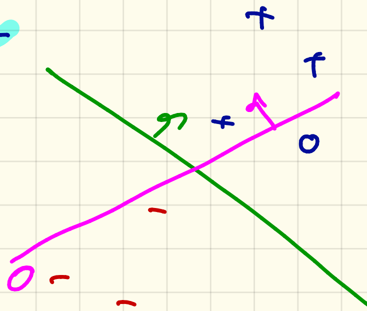
- can be used
- will converge to global min
- but will penalize well classified pts
  - ↳ decision boundary near all pts

eh?

differentiable

incorrectly classified

correctly classified



16. Consider a set of 1-dimensional data points

$$(x_1 = 0, y_1 = +1)\ (x_2 = 1, y_1 = -1)\ (x_3 = 2, y_1 = +1)\ (x_4 = 4, y_1 = +1)$$
$$(x_5 = 6, y_1 = -1)\ (x_6 = 7, y_1 = -1)\ (x_7 = 8, y_1 = +1)\ (x_8 = 9, y_1 = -1)$$

Predict **-1** or **+1** using a kNN (k-nearest neighbor) classifier with $k = 3$ on the following queries.

(a) $x = 3$ +1

(b) $x = 9$ -1

(c) $x = -1$ +1