L27: Review

Final Exam

Mon, Apr 27

3:30- 5:30p

Apr 20, 2026

FoDA

Jeff M. Phillips

supervised

unsupervised

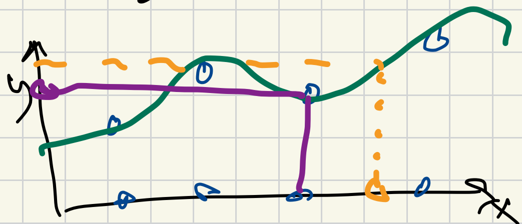
labels

Input (x, y)

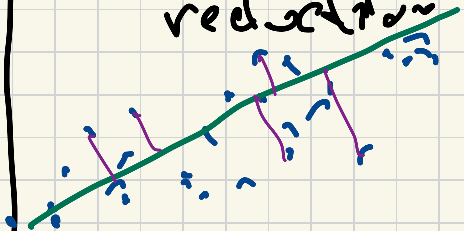
Input $X \subset \mathbb{R}^d$

output real value

regression

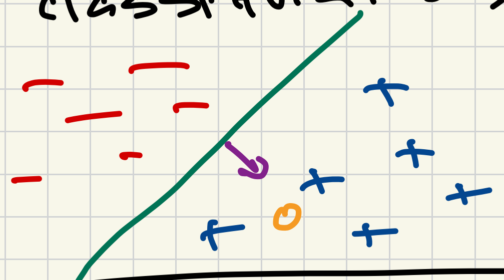


dimensionality reduction

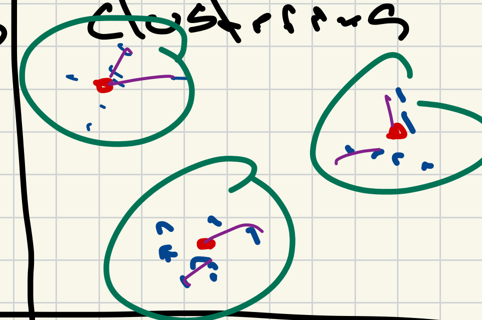


output set

classification



clustering



Split

Train

Build Models

Testing

Evaluate models

cross-validation

Sum of Squared Errors

Data $X \in \mathbb{R}^d$, model $M : \mathbb{R}^d \to \mathbb{R}^k$

$$SSE(X, M) = \sum_{x \in X} (M(x_i) - y_i)^2$$ supervi

$$\text{o.s. } \sum_{x \in X} (M(x_i) - x_i)^2$$ unsupervi

- $X \in \mathbb{R} \to$ opt mean
- $X \in \mathbb{R}^d \quad y_i \implies \alpha = (X^T X)^{-1} X^T y$
- gradient descent, strongly convex
- Power Method (for DR)
- Lloyd's (k-means)

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

$\rightarrow 0.25$

$\frac{0.20}{0.3+0.2+0.25} = \frac{0.2}{0.75}$

0.3

Derive the following values

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

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

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

Compute the following probability distributions.

$X=1$	$X=2$	$X=3$
0.4	0.3	0.3

(d) What is the marginal distribution for X?

(e) What is the conditional probability for Y, given that $X = 2$? $Pr(Y=1) = \frac{0.05}{0.3}$ $Pr(Y=2) = \frac{0.25}{0.3}$

Answer the following question about the joint distribution.

(f) Are random variables X and Y independent? $\frac{0.1}{0.4} \neq \frac{0.05}{0.3} \Rightarrow$ No.

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

0.4 $Pr(X=1) = Pr(X=1 \mid Y=1) = \frac{0.1}{0.1+0.05+0.1} = \frac{0.1}{0.25} = 0.4$

and $Pr(Y=1) = Pr(Y=1 \mid X=1)$

2. Consider two models M_1 and M_2, where from prior knowledge we believe that $\boxed{\mathbf{Pr}(M_1) = 0.25}$ and $\boxed{\mathbf{Pr}(M_2) = 0.75}$. 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?

$$Pr(M_1 \mid D) = \frac{Pr(D \mid M_1) \cdot Pr(M_1)}{Pr(D)} = \frac{(0.25)(0.5)}{Pr(D)}$$

$$Pr(M_2 \mid D) = \frac{Pr(D \mid M_2) \cdot Pr(M_2)}{Pr(D)} = \frac{(0.75)(0.01)}{Pr(D)}$$

bigger

M_1 more likely.

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

iid = independent and identically distributed

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

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

$$= \frac{1}{100}$$

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 $\mathbf{E}[\bar{X}]$? $= 3 = E[X] = 3$

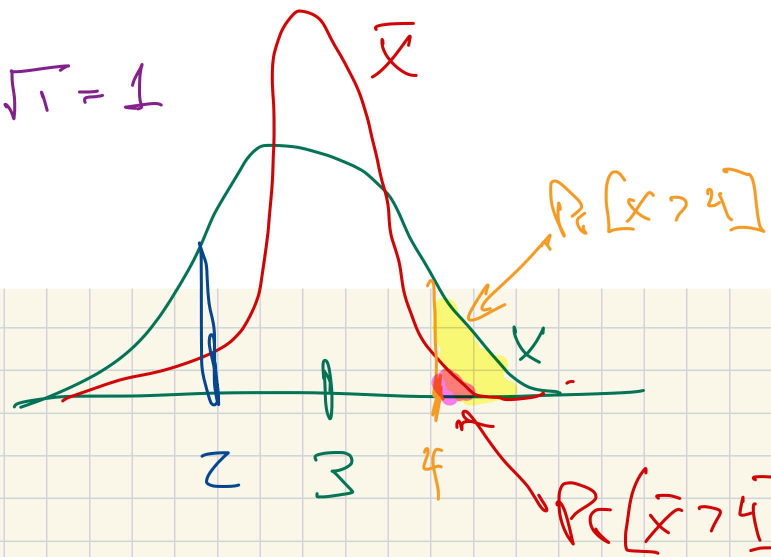
(b) What is $\mathbf{Var}[\bar{X}]$? $= \frac{Var[X]}{n} = 1$

(c) What is the standard deviation of $\bar{X}$? $= \sqrt{1} = 1$

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

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

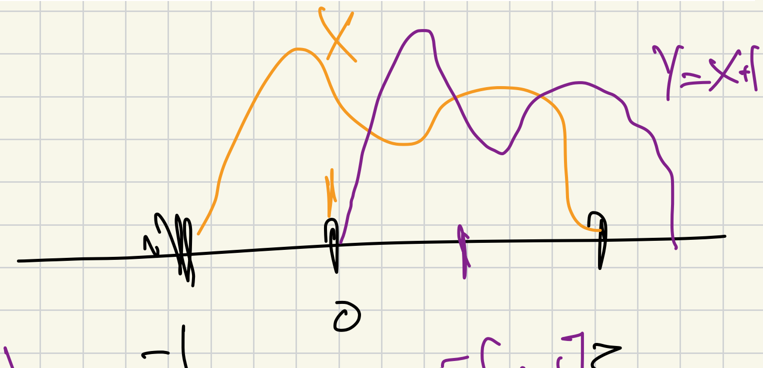
Which is <u>not</u> smaller



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.)

Markov

RV ≥ 0



$Y = X + 1$

$E[Y] = 1$

$$\Pr[Y > 2.5] \leq \frac{E[Y]}{2.5} = \frac{1}{2.5}$$

$$\Pr[X > 1.5] = \Pr[Y > 2.5] \leq \boxed{\frac{1}{2.5}}$$

6. Consider a matrix

$$A = \begin{bmatrix} \cancel{2} & \cancel{2} & \cancel{3} \\ -2 & 7 & 4 \\ -3 & -3 & -4.5 \\ -8 & 2 & 3 \end{bmatrix} \begin{bmatrix} \pi \\ e \\ i \\ 1 \end{bmatrix}$$

$\boxed{\begin{array}{l} \circ \text{ Square} \\ \circ \text{ full rank} \end{array}}$

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

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

$A = 4 \times 3$

(c) Is AA^T invertable? $[4 \times \cancel{3}][\cancel{3} \times 4] \Rightarrow [4 \times 4]$ No

(d) Is A^TA invertable? $[3 \times 4][4 \times 3] \Rightarrow [3 \times 3]$ Yes

typically $A \in \mathbb{R}^{n \times d}$

n # data points

d = dimension

$n >> d$.

$\hookrightarrow A^TA \in \mathbb{R}^{d \times d}$

$\alpha = (A^TA)^{-1}A^Ty$

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

(a) Check if u or $\cancel{v}$ is a unit vector.

$\|u\| = 1$

$\|u\|_2 = \sqrt{\sum_i u_i^2}$

$\times 2 \quad 0.25 = 0.5^2$

$\times 3 \quad 0.16 = 0.4^2$

$\times 2 \quad 0.01 = 0.1^2$

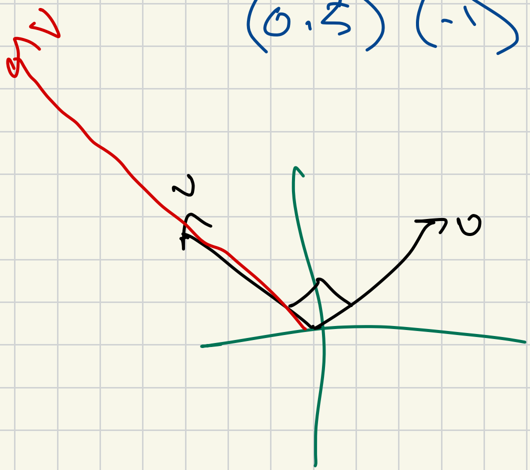
1.00

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

(c) Are u and v orthogonal? $\boxed{\text{Yes}}$ if $\langle u, v \rangle = 0$

$\boxed{\langle u, v \rangle} = \sum_i u_i v_i$

$(0.5)(-1) + (0.4)(-2) + \dots + = \boxed{0}$



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. reasonable

(b) Use k-means clustering to group the customers into 4 types using Euclidean distance between rows as the distance. not reasonable.

(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 least projection error. not reasonable

(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. reasonable

Euclidean mix units

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 $\alpha \in \mathbb{R}^3$. The remaining two testing data points are (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?

prediction on x_1 is $\langle \alpha, x_1 \rangle$

sq-error on x_1, y_1 is $(\langle \alpha, x_1 \rangle - y_1)^2$

expected $\boxed{\frac{1}{2}\left(\underbrace{(\langle \alpha, x_1 \rangle - y_1)^2}_{\text{sq err}} + \underbrace{(\langle \alpha, x_2 \rangle - y_2)^2}_{\text{sq error}}\right)}$

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.

$$v = (v_x, v_y) \implies \boxed{(v_x, v_y) - \gamma \nabla (x, y)} \quad \text{GD}$$

$$(1, 2) - 1 \begin{pmatrix} x - 1, & 2y + x \\ 1 - 1, & 2(2) + 1 \end{pmatrix}$$

$$(1, 2) - (0, 5)$$

$$(1, -3)$$

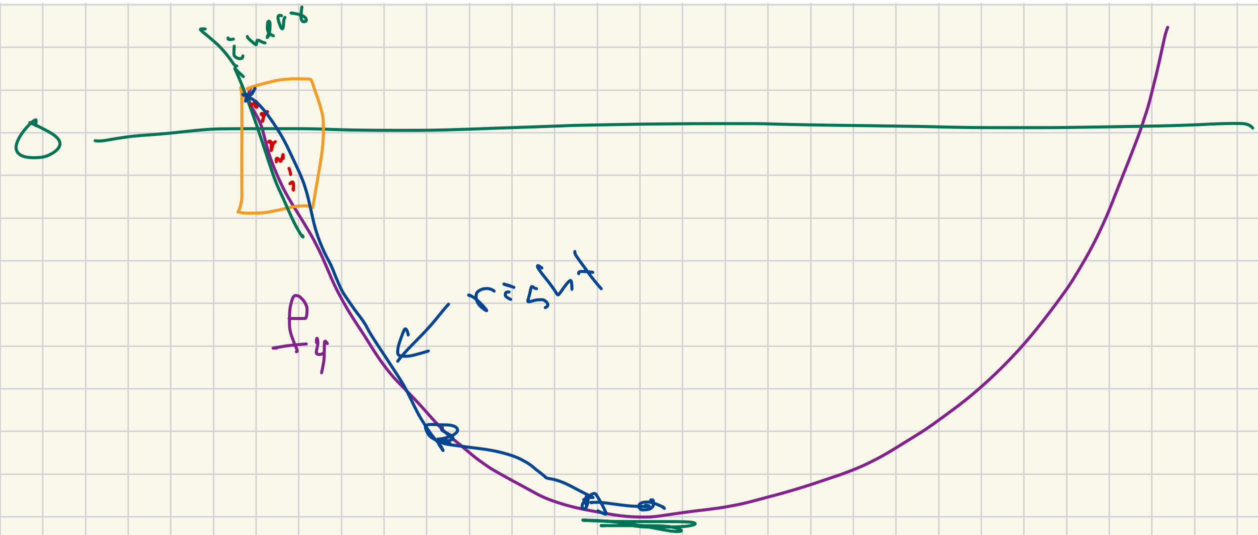
11. Consider running gradient descent with a fixed learning rate γ. For each of the following, we plot the function value over 10 steps (the function is different each time). Decide whether the learning rate is probably **too high**, **too low**, or **about right**.

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

(b) f_2: 100, 50, 75, 60, 65, 45, 75, 110, 90, 85 — 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 $\boxed{\sigma_1^2 = 10}$, $\boxed{\sigma_2^2 = 5}$, $\sigma_3^2 = 2$, and $\sigma_4^2 = 1$.

(a) What is the rank of A? $= 4$

(b) What is $\|A - \boxed{A_2}\|_F^2$, where A_2 is the best rank-2 approximation of A. $= \sigma_3^2 + \sigma_4^2 = 3$

(c) What is $\|A - A_2\|_2^2$, where A_2 is the best rank-2 approximation of A. $= \sigma_3^2$

(d) What is $\|A\|_2^2$? $= 10 = \sigma_1^2$

(e) What is $\|A\|_F^2$? $= 10 + 5 + 2 + 1$

$\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2$

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

(f) What is $\|Av_2\|^2$? $= \sigma_2^2 = 5$

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

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

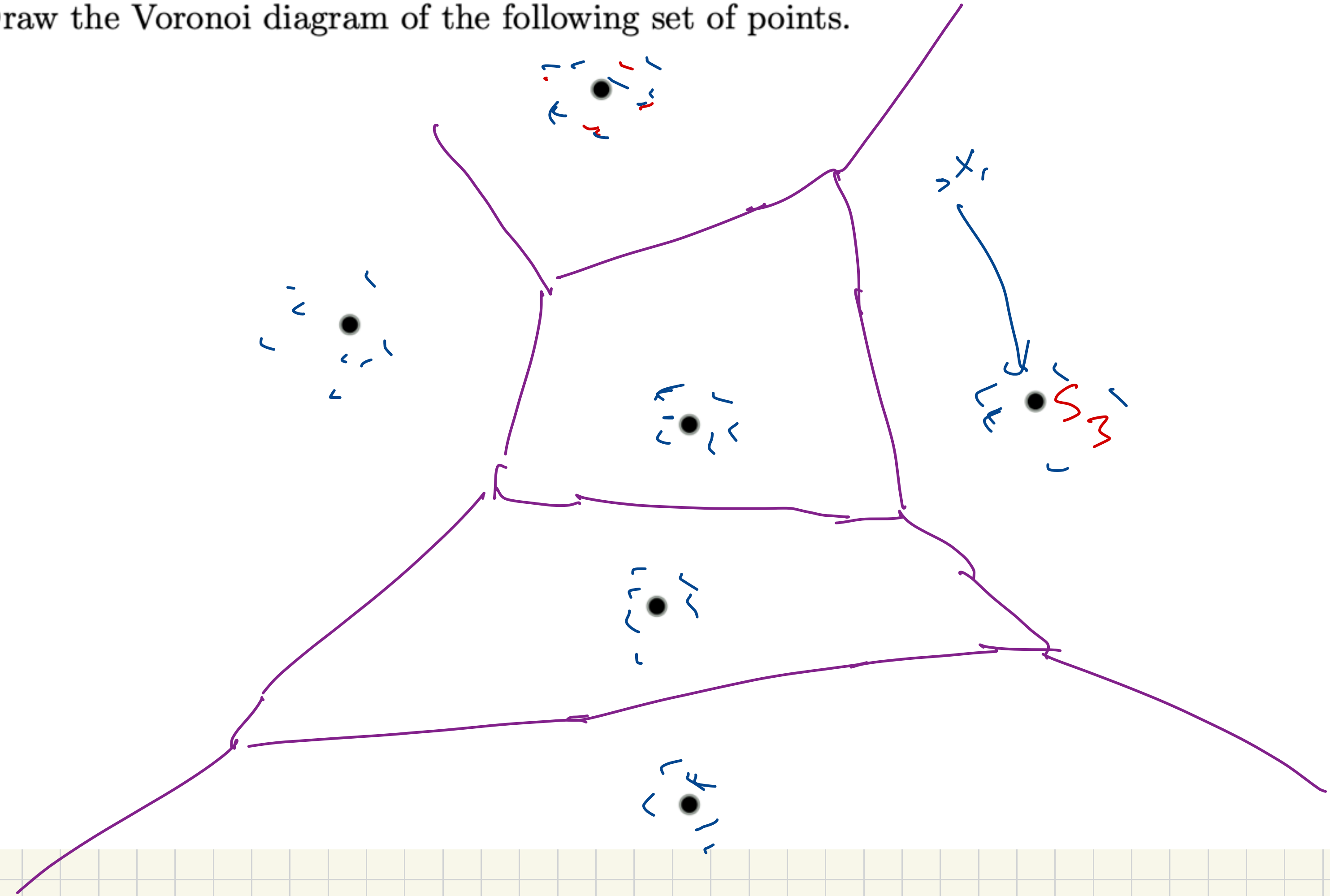
Let $a_1 \in \mathbb{R}^4$ be the first row of A.

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

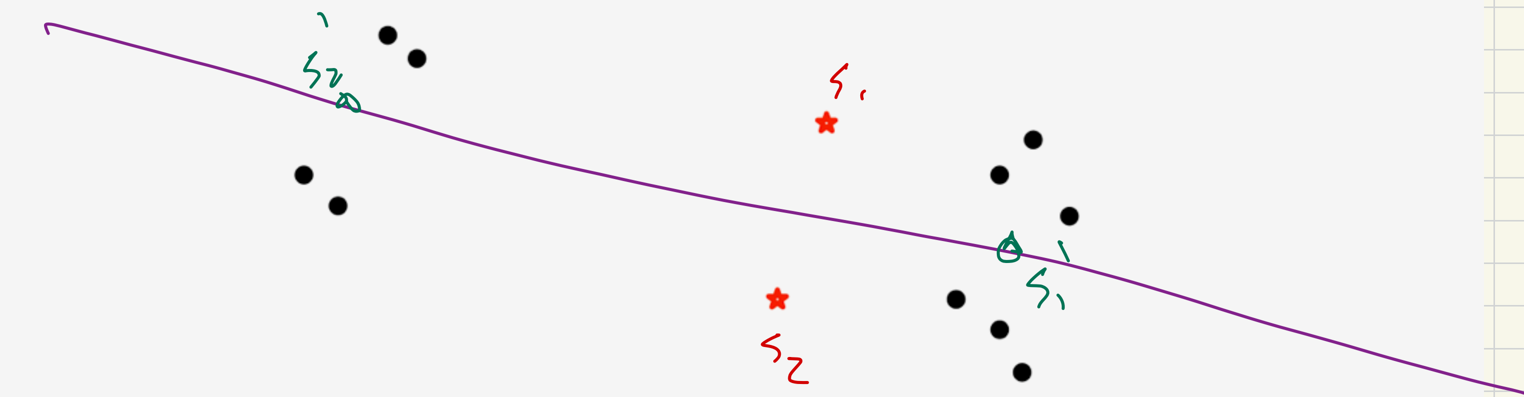
$(\langle v_1, a_1\rangle, \langle v_2, a_1\rangle, \langle v_3, a_1\rangle, \langle v_4, a_1\rangle) \in \mathbb{R}^4$

dim reduction $\in \mathbb{R}^2$

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

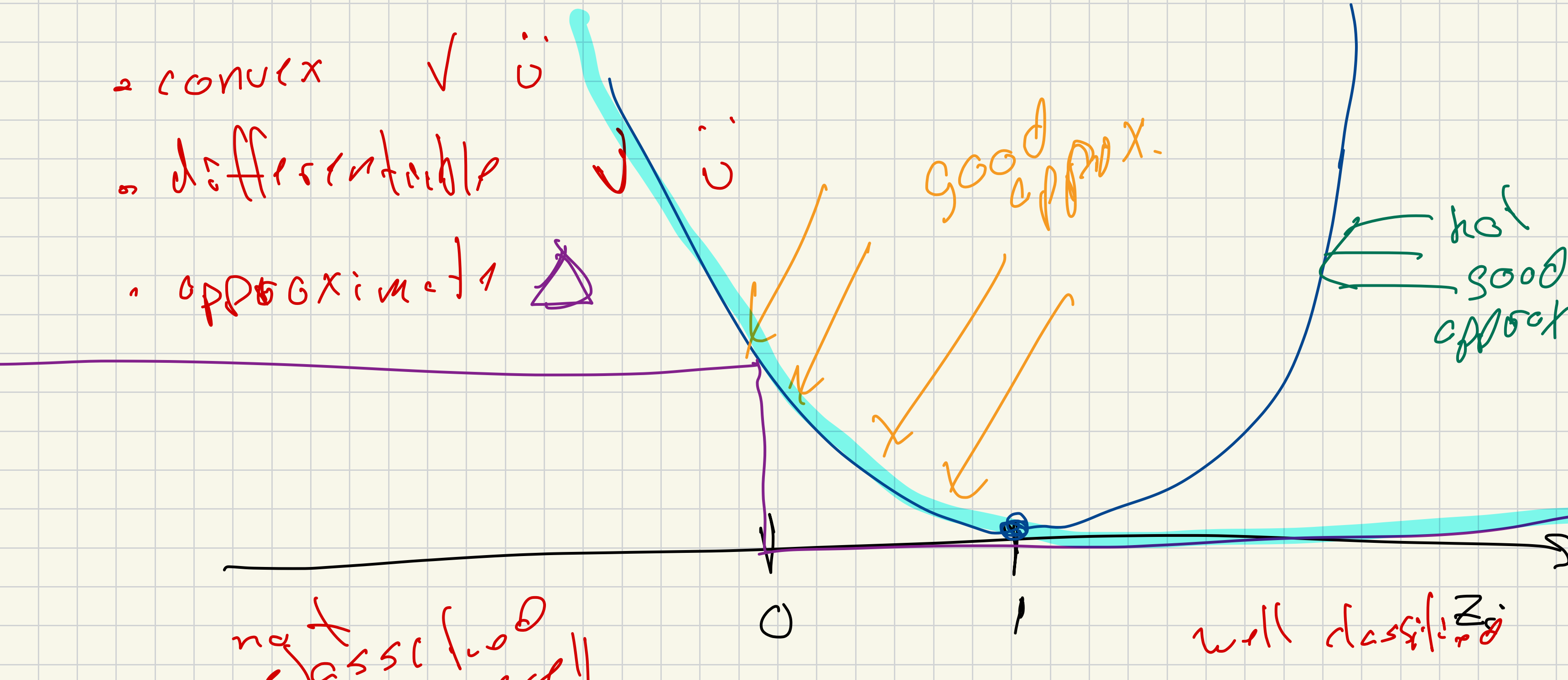


14. What should you do, if running Lloyd's algorithm for k-means clustering ($k = 2$), and you reach this scenario, where the algorithm terminates? (The black circles are data points and red stars are the centers).



Randomly restart

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.



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) $x = 3$

(b) $x = 9$

(c) $x = -1$

+1

2 +1
1 -1
→ +1

-1