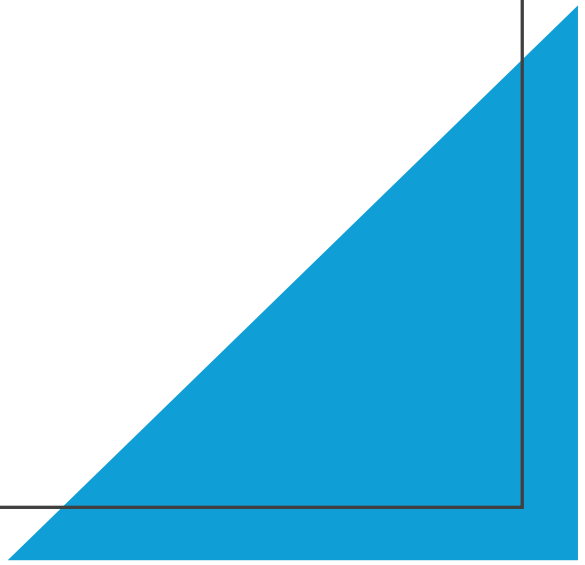
Week 2 Discussion

CSCI 567 Fall 2025



1. MULTIPLE-CHOICE QUESTIONS: One or more correct choice(s) for each question.

1.1. Which one of these is a sign of overfitting?

a. Low training error, low test error

b. Low training error, high test error

c. High training error, low test error

d. High training error, high test error

1. MULTIPLE-CHOICE QUESTIONS: One or more correct choice(s) for each question.

1.1. Which one of these is a sign of overfitting?

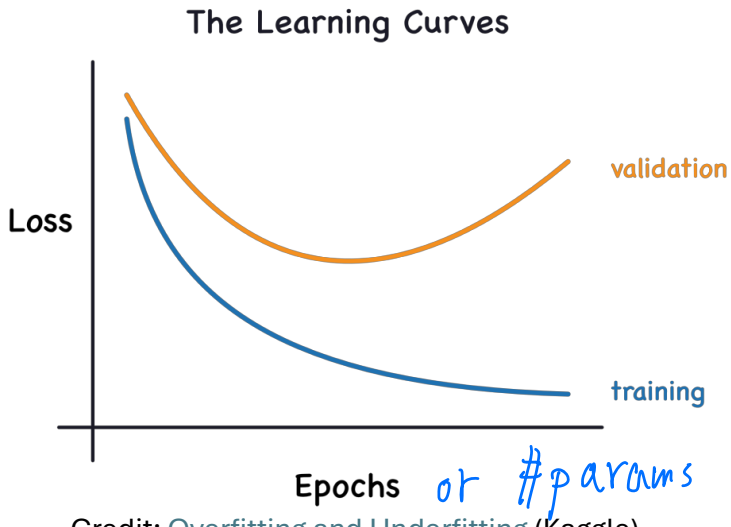
a. Low training error, low test error

b. Low training error, high test error

c. High training error, low test error

d. High training error, high test error

This is like the "definition" of overfitting



Credit: Overfitting and Underfitting (Kaggle)

1.2. Which of the following can help prevent overfitting?

a. Using more training data

b. Training until you get the smallest training error

c. Including a regularization term in the loss function

d. All of the above

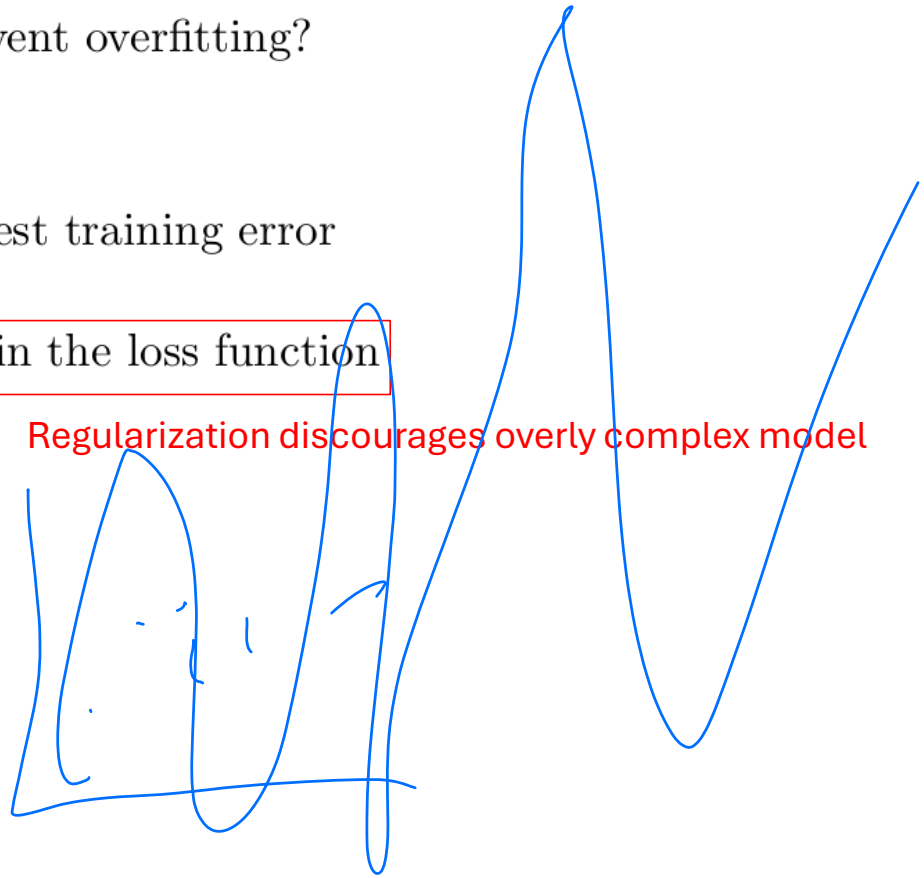
1.2. Which of the following can help prevent overfitting?

a. Using more training data

b. Training until you get the smallest training error

c. Including a regularization term in the loss function

d. All of the above



1.3. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be a data matrix with each row corresponding to the feature of an example and $\mathbf{y} \in \mathbb{R}^{N}$ be a vector of all the outcomes. The least square solution is $(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$. Which of the following is the least square solution if we scale each data point by a factor of 4 (i.e. the new dataset is $4\mathbf{X}$)?

a. $4(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

b. $\frac{1}{4}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

c. $\frac{1}{2}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

d. None of the above

1.3. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be a data matrix with each row corresponding to the feature of an example and $\mathbf{y} \in \mathbb{R}^{N}$ be a vector of all the outcomes. The least square solution is $(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$. Which of the following is the least square solution if we scale each data point by a factor of 4 (i.e. the new dataset is $4\mathbf{X}$)?

a. $4(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

b. $\frac{1}{4}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

c. $\frac{1}{2}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$

d. None of the above

$$(cX)^T = c \cdot X^T$$

$$(cX)^{-1} = \frac{1}{c} \cdot X^{-1}$$

$$= \left((4X)^T (4X) \right)^{-1}$$

$$= (16 X^T X)^{-1}$$

$$= \frac{1}{16} X^{-1} X^{-T} \cdot 4 X^T$$

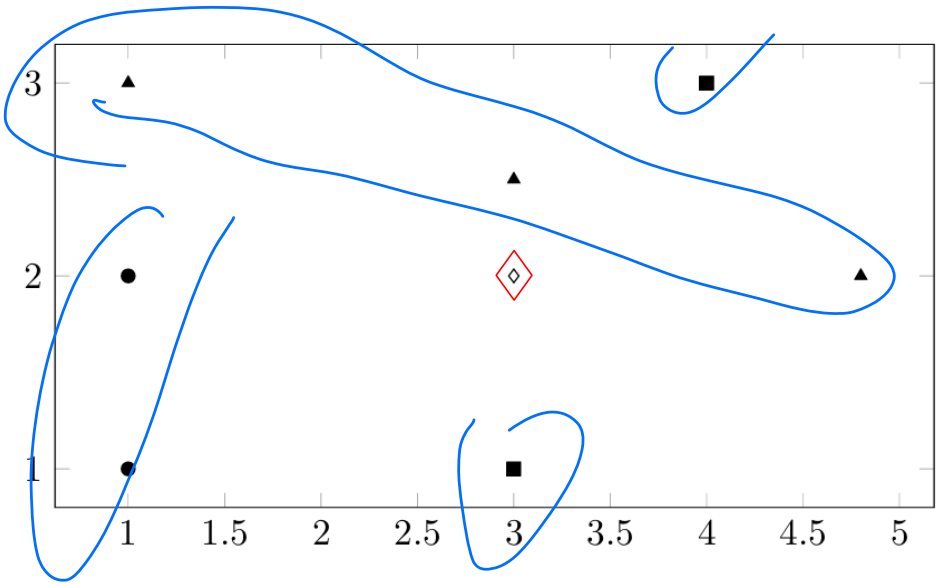
1.4. Consider the following two-dimensional dataset with $N = 7$ training points of three classes (triangle, square, and circle), and additionally one test point denoted by the diamond. Which of the following configuration of the K-nearest neighbor algorithm will predict triangle for the test point?

a. $K = 1$, L2 distance

b. $K = 3$, L1 distance

c. $K = 3$, L2 distance

d. $K = 7$, any distance



$\bar{x} \in \mathbb{R}^D$, $x_i \in \mathbb{R}^D$, $i = 1, \ldots, N$

$\|\bar{x} - x_i\|_1 = \sum_d |\bar{x}_d - x_{id}|$

$\|\bar{x} - x_i\|_2 = \sqrt{\sum_d (\bar{x}_d - x_{id})^2}$

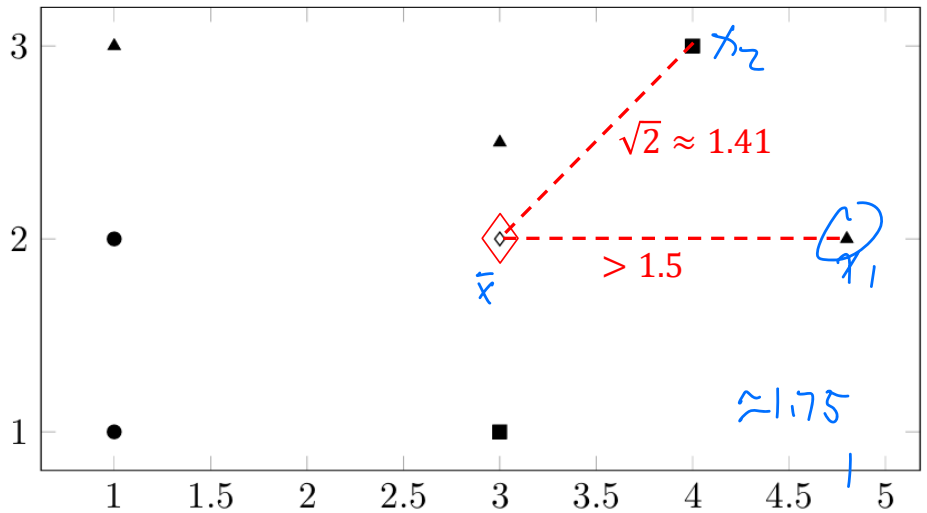
1.4. Consider the following two-dimensional dataset with $N = 7$ training points of three classes (triangle, square, and circle), and additionally one test point denoted by the diamond. Which of the following configuration of the K-nearest neighbor algorithm will predict triangle for the test point?

a. $K = 1$, L2 distance

b. $K = 3$, L1 distance

c. $K = 3$, L2 distance

d. $K = 7$, any distance



$\|\bar{x} - x_2\|_1 = 2$ $\|\bar{x} - x_1\|_1 \simeq 1.75$

1.5. Which of the following on linear regression is correct?

a. The least square solution has a closed-form formula, even if L2 regularization is applied.

b. The covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not invertible if and only if the number of data points N is smaller than the dimension D.

c. When the covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not invertible, the Residual Sum of Squares (RSS) objective has no minimizers.

d. Linear regression is a parametric method.

1.5. Which of the following on linear regression is correct?

a. The least square solution has a closed-form formula, even if L2 regularization is applied.

$w^* = (X^T X + \lambda \Sigma)^{-1} X^T y$.

b. The covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not invertible if and only if the number of data points N is smaller than the dimension D.

Even if $N \geq D$, $X^T X$ can still be **not** invertible , e.g., due to "collinearity"

c. When the covariance matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not invertible, the Residual Sum of Squares (RSS) objective has no minimizers.

Infinitely many solutions

d. Linear regression is a parametric method.

2. Nearest Neighbor Classification

We mentioned that the Euclidean/L2 distance is often used as the *default* distance for nearest neighbor classification. It is defined as

$$E(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||_2 = \sqrt{\sum_{d=1}^{D} (x_d - x'_d)^2} \tag{1}$$

In some applications such as information retrieval, the cosine distance is widely used too. It is defined as

$$C(\mathbf{x}, \mathbf{x}') = 1 - \frac{\mathbf{x}^\intercal \mathbf{x}'}{||\mathbf{x}||_2 ||\mathbf{x}'||_2} = 1 - \frac{\sum_{d=1}^{D} (x_d \cdot x'_d)}{||\mathbf{x}||_2 ||\mathbf{x}'||_2} \tag{2}$$

where the L2 norm of $\mathbf{x}$ is defined as

$$||\mathbf{x}||_2 = \sqrt{\sum_{d=1}^{D} x_d^2}. \tag{3}$$

Show that, if data is normalized with unit L2 norm, that is, $||\mathbf{x}||_2 = 1$ for all $\mathbf{x}$ in the training and test sets, changing the distance function from the Euclidean distance to the cosine distance will NOT affect the nearest neighbor classification results.

$x^T x = \|x\|_2^2$

When $\|x\| = \|x'\| = 1$,

$C(x, x') = 1 - x^T x'$

$E(x, x')^2 = (x - x')^T (x - x') = \|x\|^2 + \|x'\|^2 - 2 x^T x' = 2\, C(x, x')$.

($\|x\| = 1$, $\|x'\| = 1$)

3. Linear Regression

In the lectures, we have described the least mean square solution for linear regression as

$$\boldsymbol{w}^* = (\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{y}$$

where $\tilde{\boldsymbol{X}}$ is the design matrix (N rows, $D+1$ columns) and $\boldsymbol{y}$ is the N-dimensional column vector of the true values in the training data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$.

Question 3.1 We mentioned a practical challenge for linear regression: when $\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$ is not invertible. Please use a concise mathematical statement (*in one sentence*) to summarize the relationship between the training data $\tilde{\boldsymbol{X}}$ and the dimensionality of $\boldsymbol{w}$ when this scenario happens. Then use this statement to explain why this scenario must happen when $N < D+1$.

Why? If $N < D+1$, then $\tilde{X}^T\tilde{X}$ not invertible.

3. Linear Regression

In the lectures, we have described the least mean square solution for linear regression as

$$\boldsymbol{w}^* = (\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{y}$$

where $\tilde{\boldsymbol{X}}$ is the design matrix (N rows, $D+1$ columns) and $\boldsymbol{y}$ is the N-dimensional column vector of the true values in the training data $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$.

Question 3.1 We mentioned a practical challenge for linear regression: when $\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}$ is not invertible. Please use a concise mathematical statement (*in one sentence*) to summarize the relationship between the training data $\tilde{\boldsymbol{X}}$ and the dimensionality of $\boldsymbol{w}$ when this scenario happens. Then use this statement to explain why this scenario must happen when $N < D+1$.

$N < D + 1 \Rightarrow \operatorname{rank}(\tilde{X}^T\tilde{X}) < D + 1$

$\Leftrightarrow \tilde{X}^T\tilde{X}$ not invertible

Why?

- $\tilde{X}^T\tilde{X}$ is $(D+1) \times (D+1)$ matrix
- $\operatorname{rank}(\tilde{X}^T\tilde{X}) = \operatorname{rank}(\tilde{X}) \leq \min\{N, D+1\}$

In general, invertible $\Leftrightarrow$ full rank

Question 3.2 In this problem we use the notation $w_0 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ for the linear model, that is, we do not append the constant feature 1 to x. In the lecture we saw that when $D = 0$, the bias w_0^* is simply the mean of the sample responses

$$w_0^* = \frac{1}{N}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{y} = \frac{1}{N}\sum_n y_n, \tag{4}$$

where $\mathbf{1}_N = [1, 1, \ldots, 1]^{\mathrm{T}}$ is an N-dimensional column vector whose entries are all ones. Now, we would like you to generalize this to arbitrary D and arrive at a more general condition where Eqn. (4) holds.

Design matrix *(without bias)*

$$X = \begin{pmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_N^T & - \end{pmatrix} \in \mathbb{R}^{N\times D}$$

Prediction

$\hat{y}_i = x_i^T w + w_0$ $\quad \hat{y} = Xw + w_0 1_N$

"Residual" error

$\epsilon_i = y_i - \hat{y}_i = y_i - x_i^T w - w_0$ $\quad \epsilon = y - \hat{y} = y - Xw - w_0 1_N$

objective minimize $(w_0, w) \in \mathbb{R}^{D+1}$ RSS.

1) write down the residual sum of squares objective w.r.t. the variable of interest;

Fix any w^*,

$$w_0^* = \operatorname*{argmin}_{w_0} \|y - Xw^* - w_0 1_N\|_2^2$$

2) take derivative with respect to w_0 and set it to 0;

$$0 = -2\, 1_N^T (y - Xw^* - w_0^* 1_N) \Rightarrow N w_0 = 1_N^T y - (1_N^T X) w^*$$

$(1_N^T X) = 0_D^T$

$$\sum_{i=1}^N \epsilon_i^2 = \|\epsilon\|_2^2 = \epsilon^T \epsilon.$$

$$RSS = \|y - Xw - w_0 1_N\|_2^2$$

3) solve the obtained equation and conclude that Eqn. (4) holds if

$$\frac{1}{N}\sum_n x_{nd} = 0, \quad \forall d = 1, 2, \ldots, D, \tag{5}$$

(5) $\Rightarrow \sum_n x_{nd} = 0 \Rightarrow [X^T 1_N]_d = 0$

that is, each feature has zero mean.

$$w_0 = \frac{1}{N} 1_N^T y = \frac{1}{N}\sum_i y_i.$$

$$\frac{\partial \epsilon_i}{\partial w_0} = -1.$$

$$\frac{\partial \|\epsilon\|_2^2}{\partial w_0} = \frac{\partial \epsilon_i}{\partial w_0} = -2\epsilon_i$$

$$\frac{\partial \|\epsilon\|^2}{\partial w_0} = -2\sum_i \epsilon_i = -2 \cdot 1_N^T \epsilon = 0$$