Week 2 Practice

CSCI 567 Machine Learning

Spring 2025

Instructor: Haipeng Luo

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

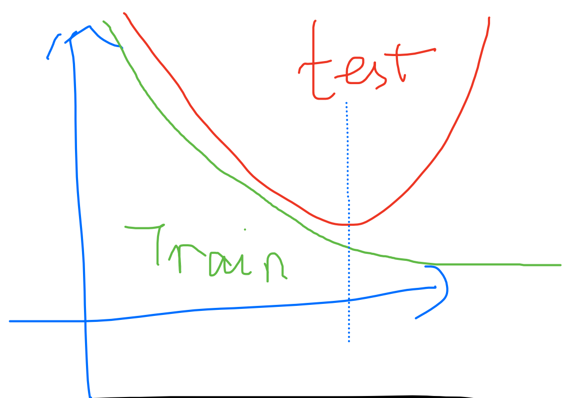
   1.1. Which one of these is a sign of overfitting?

   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



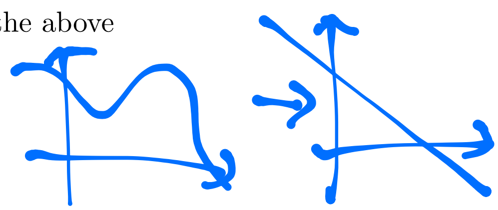
   1.2. Which of the following can help prevent overfitting?

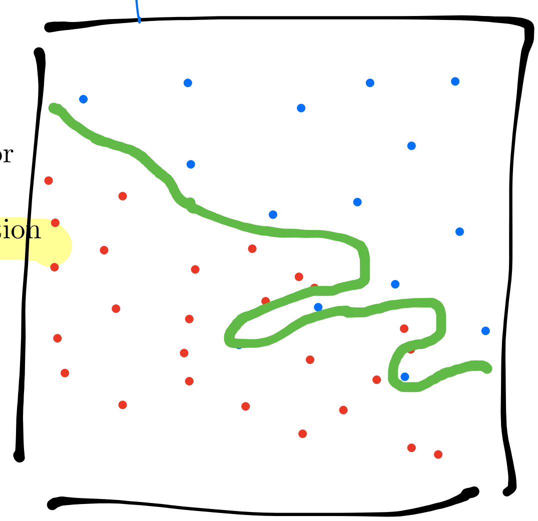
   a. Using more training data

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

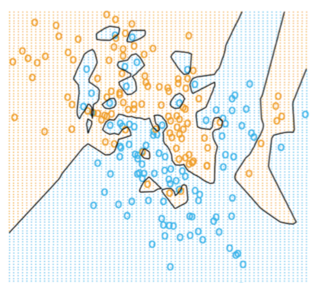
d. None of the above

$$(4X^T)(4X)w - (4X)^T y = 0$$

$$16X^T Xw - 4X^T y = 0$$

$$= \frac{4}{16}(X^TX)^{-1}X^Ty = \frac{1}{4}(X^TX)^{-1}X^Ty$$

1.4. Which of these classifiers could have generated this decision boundary?



a. Regularized Linear Regression

b. Regression with non-linear basis

low dimensional

c. 1-nearest-neighbor

d. None of the above

2. Nearest Neighbor Classification

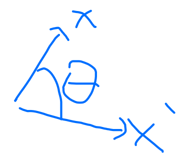
We mentioned that the Euclidean/L2 distance is often used as the *default* distance for nearest neighbor classification. It is defined as

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

cosine similarity

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

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



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

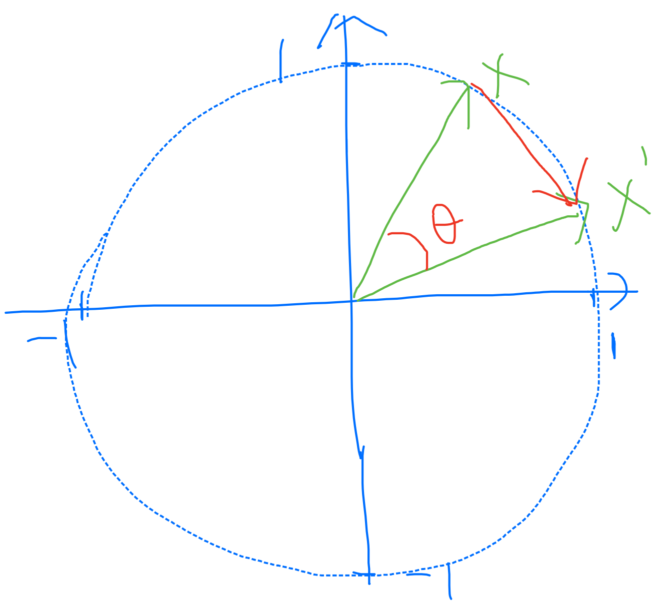
$$||\mathbf{x}||_2 = \sqrt{\sum_{d=1}^{D} x_d^2}. \tag{3}$$

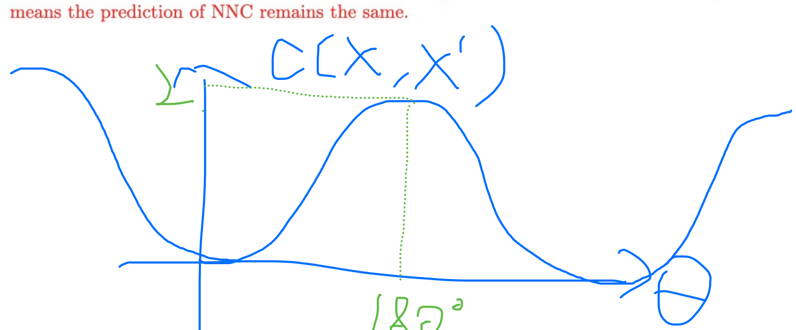
$\cos\theta \, ||x||_2 \, ||x'||_2 = x^T \cdot x'$

$\cos\theta = \frac{x^T x'}{||x||_2 ||x'||_2}$

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.

When data is normalized, we have $E(\mathbf{x}, \mathbf{x}')^2 = \sum_{d=1}^{D} (x_d - x'_d)^2 = \sum_{d=1}^{D} (x_d^2 + {x'_d}^2 - 2x_d x'_d) = 2(1 - \sum_{d=1}^{D} (x_d \cdot x'_d)) = 2C(\mathbf{x}, \mathbf{x}')$. Therefore, the nearest neighbor of a point in terms of the cosine distance is exactly the same as its nearest neighbor in terms of the L2 distance, which means the prediction of NNC remains the same.





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 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 $N < D+1$.

$r(\tilde{\boldsymbol{X}}) < D+1$ where $r(\boldsymbol{M})$ is the rank of matrix $\boldsymbol{M}$. Since $r(\tilde{\boldsymbol{X}}) \le \min\{N, D+1\}$, it must be smaller than $D+1$ when $N < D+1$.

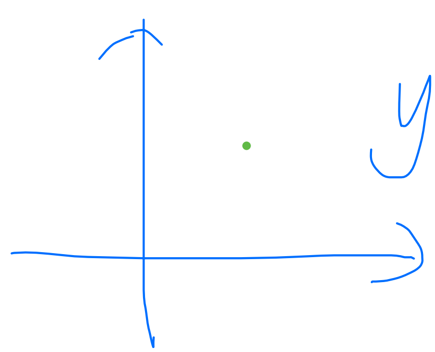
Rank of Matrix M is the dimension of the lesser of row or column space of M

$$r(\tilde{X}^T\tilde{X}) \le \min(r(\tilde{X}^T), r(x))$$
$$< N < D+1$$

$\tilde{X}^T\tilde{X}$ is $D+1 \times D+1$ matrix

not invetible

e.g.



$y = kx + b$

1 data point

Two dimensional column vector and 1 data point

Question 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. Please do so by following the three steps below:
1) write down the residual sum of squares objective w.r.t. the variable of interest;
2) take derivative with respect to w_0 and set it to 0;
3) solve the obtained equation and conclude that Eqn. (4) holds if

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

that is, each feature has zero mean.

$$w_0^* = \arg\min_{w_0} \|\boldsymbol{y} - w_0\mathbf{1}_N - \boldsymbol{X}\boldsymbol{w}^*\|^2 \qquad \text{Residual sum of squares}$$

$$\mathbf{1}_N^{\mathrm{T}}(\boldsymbol{y} - w_0^*\mathbf{1}_N - \boldsymbol{X}\boldsymbol{w}^*) = \mathbf{0} \qquad \text{Setting derivatives w.r.t. } w_0 \text{ to } \mathbf{0}$$

$$w_0^* = \frac{1}{N}(\mathbf{1}_N^{\mathrm{T}}\boldsymbol{y} - \mathbf{1}_N^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}^*) \qquad \text{solve for } w_0^*$$

$$= \frac{1}{N}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{y} \qquad \frac{1}{N}\sum_n x_{nd} = 0 \Leftrightarrow \mathbf{1}_N^{\mathrm{T}}\boldsymbol{X} = \mathbf{0}$$

Thus, if the feature values are zero on average, the bias w_0^* is the average response of training samples.

$$\frac{\partial \|y - w_0 1_N - Xw^*\|^2}{\partial w_0} \quad \text{chain rule}$$

$$= 2\, 1_N (y - w_0 1_N - Xw^*) = 0$$