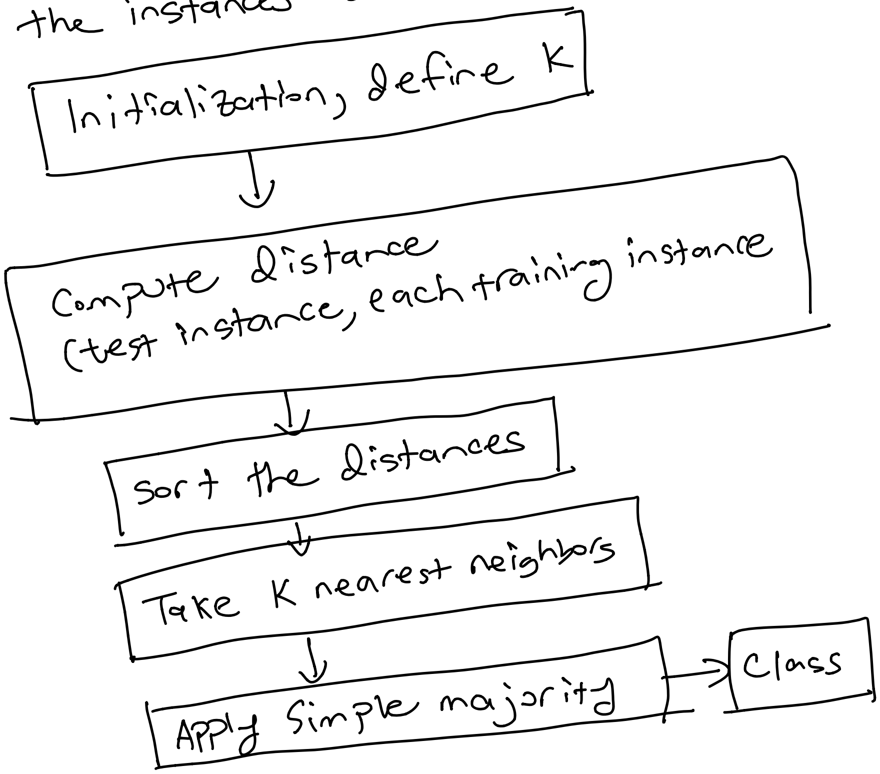
Instance based classification:

The main idea behind instance-based classification is that similar instances have similar classification.

One of the simplest examples of instance based classification is K-nearest neighbor (K-NN) classifier. In K-NN an instance is assigned to the most common class among the instances similar to it.



One of the main drawbacks of K-NN is the curse of dimensionality. As the feature space gets larger, the feature vectors become sparser and as a result distance between them increases. and distance can be dominated by irrelevant attributes. As a result, the neighborhood of a test instance doesn't contain instances that are 'similar' to it.

Practice problems on K-NN:

We are given a test point x and it's K nearest neighbors $\{z_1, z_2, \ldots, z_K\}$

(a) Let E_1 be the event that 1-NN classifier makes a mistake on X. So,

$$P(E_1)$$
$$= P(\text{label}(z_1) \neq \text{label}(x))$$
$$= 0.1$$

(b) Let E_3 be the event that 3-NN classifier makes a mistake on X

E_3 occurs when at least 2 of the 3 nearest neighbors of X have a label that is different from the label of X.

Z_1	Z_2	Z_3	Prob
D	D	D	$0.1 \times (0.2)^2$
D	D	S	$0.1 \times 0.2 \times 0.8$
D	S	D	$0.1 \times 0.8 \times 0.2$
S	D	D	$0.9 \times (0.2)^2$

$\therefore P(E_3) = 0.004 + 0.016 + 0.016 + 0.036$
$= 0.072$

(c) since $P(E_3) < P(E_1)$ so 3-NN classifier is more robust than 1-NN classifier.

2

(a) From the plot we can see that $K=5$ and $K=11$ has the lowest mean validation error but $K=5$ has a smaller variance and hence better generalization and is more stable. $\therefore k^* = 5$

(b) As K increases the variance in the 5-fold CV error increases.

Linear classification:

It is a more powerful and systematic approach to classification. It has two major components:

(a) A score function that maps the raw data to class scores

(b) A loss function that measures the goodness of the scoring function in predicting the labels.

In linear classification, we use a linear scoring function

$$f(x^{(i)}, W, b) = Wx^{(i)} + b$$

where,

$$W = \begin{bmatrix} - w_1^T - \\ - w_2^T - \\ \vdots \\ - w_C^T - \end{bmatrix}$$

where C is the number of classes. The j^{th} entry of $f(x^{(i)}, W, b)$ is the confidence score of image $x^{(i)}$ belonging to class j. W and b are parameters of the scoring function.

Softmax classifier:

In softmax classifier we view the scores as unnormalized log probabilities for each class and use a cross-entropy loss of the form

$$L_i(\theta) = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_{y_j}}}\right)$$

where $f_{y_i} = w_i^T x^{(i)} + b_i$

$f_{y_j} = w_j^T x^{(j)} + b_j$

Then the loss for softmax is

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} L_i(\theta)$$

Practice problem on softmax classifier:

$$f(z_j) = \frac{e^{z_j}}{\sum_K e^{z_K}}$$

Recall the quotient rule from basic calculus

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$$

(a) Using the quotient rule and assuming $i = j$,

$$\frac{\partial}{\partial z_i}\left[\frac{e^{z_i}}{e^{z_1} + e^{z_2} + \dots + e^{z_i} + \dots + e^{z_n}}\right].$$

$$= \frac{e^{z_i} \cdot \sum_K e^{z_K} - e^{z_i} \cdot e^{z_i}}{\left(\sum_K e^{z_K}\right)^2}$$

$$= \frac{e^{z_i}}{\sum_K e^{z_K}} - \frac{(e^{z_i})^2}{(\sum_K e^{z_K})^2}$$

$$= f(z_i)(1 - f(z_i))$$

(b) Using the quotient rule and assuming $i \neq j$,

$$\frac{\partial}{\partial z_i}\left[\frac{e^{z_j}}{e^{z_1} + e^{z_2} + \dots + e^{z_i} + \dots e^{z_j} + \dots e^{z_n}}\right]$$

$$= \frac{0 \cdot \sum_K e^{z_K} - e^{z_j} \cdot e^{z_i}}{(\sum_K e^{z_K})^2}$$

$$= -\frac{e^{z_j}}{\sum_K e^{z_K}} \cdot \frac{e^{z_i}}{\sum_K e^{z_K}}$$

$$= -f(z_j)f(z_i).$$