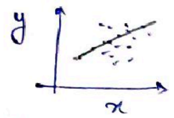
Logistic Regression

Classic Machine Learning

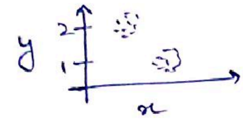
```
def sigmoid(x):
    return 1/(1+np.e**-x)
```

[e → Euler's constant]
2.718....

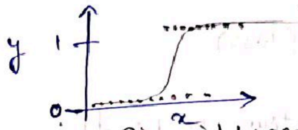
Linear Regression → predicting continous/numerical values → $y = wx + b$



Regression
↓
(finding the best fit line)



classification



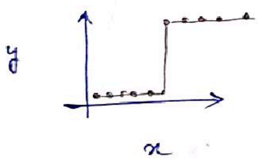
Sigmoid function

$$\sigma(y) = \frac{1}{1+e^{-y}} \quad [0 \le \sigma(y) \le 1]$$

$y = -\infty,\ \sigma(y) = 0$

$y = \infty,\ \sigma(y) = 1$

$y = 0,\ \sigma(y) = 0.5$



$$\frac{\Delta y}{\Delta x} = \frac{-}{0}$$

why not step function → Non-differentiable.

Likelihood → $p^y (1-p)^{(1-y)}$

log-likelihood → $y \log(p) + (1-y).\log(1-p)$

or

$y\ \ln(p) + (1-y).\ln(1-p)$

Negative-log-likelihood → $-\left(\sum_{i=1}^{m} y^{(i)} \log(p^i) + (1-y^i)\log(1-p^i)\right)$

↓ ensures that the loss surface has only one global minimum making optimization using GD much easier.

MSE + Sigmoid → Non-Convex

Logloss + Sigmoid → Convex

why do we not use MSE for classification? "(FB interview)"

```
def log-loss(y, y_hat):
    loss = y * np.log(y_hat) + (1-y) * np.log(1-y_hat)
    return -loss.
```

for m-points?

→ -np.mean(loss)

```
def accuracy(y_true, y_pred):
    return np.sum(y_true==y_pred)/y_true.shape[0]

accuracy(y_train, model.predict(X_train))
```

Which of the following techniques is used to address overfitting in linear regression?
a) Ridge regression.
b) Normalization.
c) Principal Component Analysis (PCA).
d) Variance Inflation Factor (VIF).

> Ans : Option (a)
>
> L2 or Ridge regularization technique that adds a penalty term to the linear regression cost function. This penalty term discourages the model from fitting the training data too closely, thereby reducing overfitting.

> **b) Normalization:** This is a data preprocessing technique used to scale numerical features to a common range, typically between 0 and 1. While normalization can improve the performance of some machine learning algorithms, it does not directly address overfitting.
>
> **c) Principal Component Analysis (PCA):** This is a dimensionality reduction technique that transforms the original features into a smaller set of uncorrelated features called principal components. While PCA can sometimes help reduce overfitting by removing redundant information, it's not a direct method for addressing overfitting.
>
> **d) Variance Inflation Factor (VIF):** This is a metric used to detect multicollinearity among independent variables in a regression model. Multicollinearity can make it difficult to interpret the coefficients and can lead to unstable estimates. While addressing multicollinearity can indirectly improve model performance, VIF itself is not a technique for directly addressing overfitting.

```
A company is building a credit scoring model to predict risk of default.
Which function can be used to map the model's output to a probability between 0 and 1?
A) Sigmoid function.
B) Linear function.
C) Step function.
D) Exponential function.
Correct Answer: A) Sigmoid function.
```

Why Sigmoid + MSE = Non-Convex Loss?

Because the sigmoid function is **non-linear**, applying MSE as a loss function produces a **non-convex surface** with **multiple peaks and valleys**. This happens because:

- When the predicted probability $\hat{y}_i$ is close to 0 or 1, MSE reacts slowly because the sigmoid function flattens at the ends.
- When $\hat{y}_i$ is around 0.5, MSE reacts more strongly because the sigmoid changes quickly in this region.

This uneven behavior creates **multiple minima**, making the loss function non-convex.

Solution: Use Log Loss (Binary Cross-Entropy Loss)

To avoid this issue, logistic regression uses **Log Loss**, which produces a **convex** loss surface:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Log Loss ensures that the loss surface has only **one global minimum**, making optimization using **Gradient Descent** much easier.

```
In logistic regression, the output of the sigmoid function is interpreted as:
a) Class probabilities
b) Raw scores
c) Error rates
d) Regression coefficients
```

```
The sigmoid function maps its input to a range between:
a) -1 and 1
b) 0 and 1
c) -∞ and ∞
d) 0 and ∞
```

```
In logistic regression, the cost function used is:
a) Mean Squared Error (MSE)
b) Mean Absolute Error (MAE)
c) Log-Loss
```

```
Start coding or generate with AI.
```

```
Which of the following is a characteristic of sigmoid activation function?
a) Linear activation
b) Non-linear activation
c) Step-wise activation
d) Exponential activation
```

Ans: option(d)

```
Which of the following is true about k-fold cross-validation?
a) Each fold serves as both training and testing data atleast once
b) Each fold serves as only training data and never as testing data
c) Each fold serves as only testing data and never as training data
d) The model is trained on the first fold and tested on the remaining folds
```

> Ans: option(a)

```
```
Which of the following is a drawback of k-fold cross-validation?
a) It requires more computational resources compared to the holdout method
b) It can lead to overfitting if the model is not properly regularized
c) It is less accurate than traditional train-test splits
d) It does not require hyperparameter tuning
```

> Ans option(a) is also the answer but correct and most optimal answer is option (b).

```
Which statement about the step function is true?
a) It is continuous and differentiable
b) It is continuous but not differentiable
c) It is neither continuous nor differentiable
d) It is differentiable but not continuous
```

> Ans: option(c)

```
What is the main risk of overfitting when tuning hyperparameters in logistic regression?
a) The model may generalize well to unseen data but poorly on the training data
b) The model may perform well on the training data but poorly on unseen data
c) The model may underperform compared to a model with default hyperparameter values
d) The model may be too simple and fail to capture complex relationships in the data
```

> Ans: option(b)
```