Machine Learning

Prof. Adil Khan

Objectives

1. A quick recap of last week
2. Gradient Descent
3. Classification Tasks
4. What is logistic regression? What is its objective function? How is it motivated?
5. Finding the optimum parameters of logistic regression
6. Classification Metrics1: Accuracy
7. Things to Know about ML-1: Class-imbalance
8. Classification Metrics2: Precision, Recall, F1-score

Recap (1)

1. What are software bugs?
2. What are their sources?
3. What are their adverse effects?
4. How unlikely is it to create bug-free software?
5. How important is it to be able to predict defect's related information?

Predicting Number of Defects From the Point of view of ML

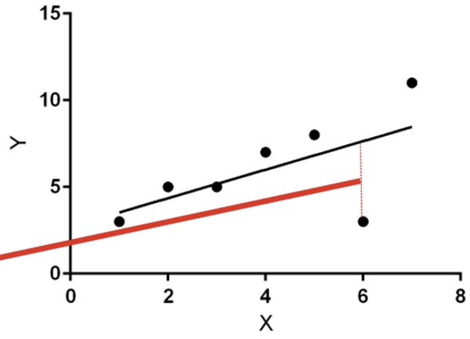
$$\underbrace{\#\ of\ defects\ in\ p_i}_{y_i} = f(\underbrace{features\ or\ behavior\ of\ p_i}_{\boldsymbol{x_i}})$$

y categorical is continuous?

Thus we are dealing with a regression problem

How Do We Train Linear Regression Model?

$$f(x_i) = w_0 + w_1 x_i$$



$$e_i = y_i - f(x_i)$$

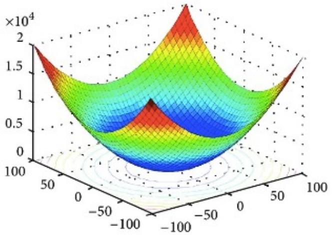
Linear Regression

$$\boxed{y = w_0 + w_1 x_1 + w_2 x_2 + \cdots w_p x_p}$$

- The response variable is quantitative
- The relationship between response and predictors is assumed to be linear in the inputs
- Thus we are restricting ourselves to a hypothesis space of linear functions

Recap (2)

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$



Objective Function

$$f(x_i) = w_0 + w_1 x_i$$

$$e_i = y_i - f(x_i)$$

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

$$\underset{w_o, w_1}{\operatorname{argmin}} \; \mathcal{L}(w_0, w_1)$$

Thus, it is convex, at the unique minimum of our loss function, its "partial" derivative with respect to w_0 and w_1 will be zero!

The Least Square Solution

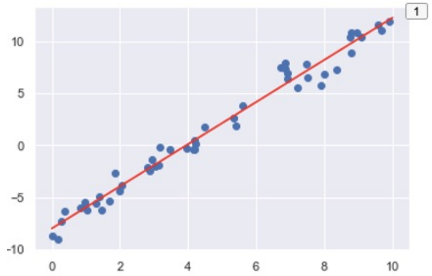
$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

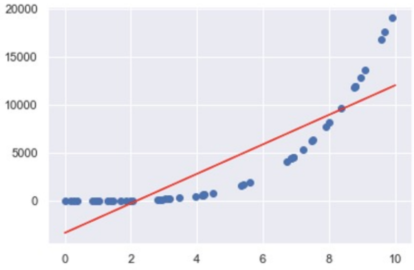
1. Compute partial derivatives of the loss function with respect to w_0 and w_1
2. Set them to 0
3. And solve for w_0 and w_1

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

Recap (3)





Polynomial Regression

- That is,

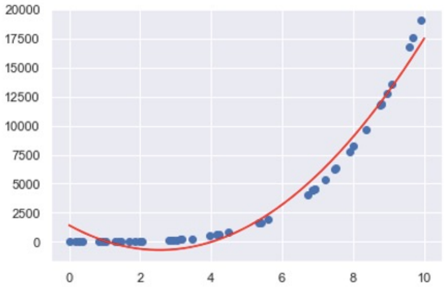
$$y = w_0 + w_1 x + w_2 x^2$$

- More generally,

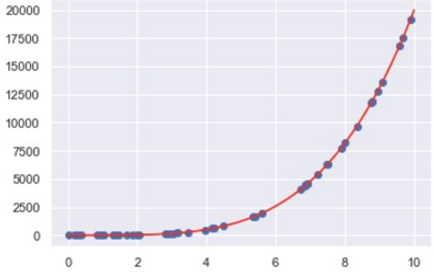
$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$

- Do not forget, “the model is still linear in parameters”

- Using the same framework that we learned, to fit a family of more complex models through a <u>transformation of predictors</u>



Order or Degree 2



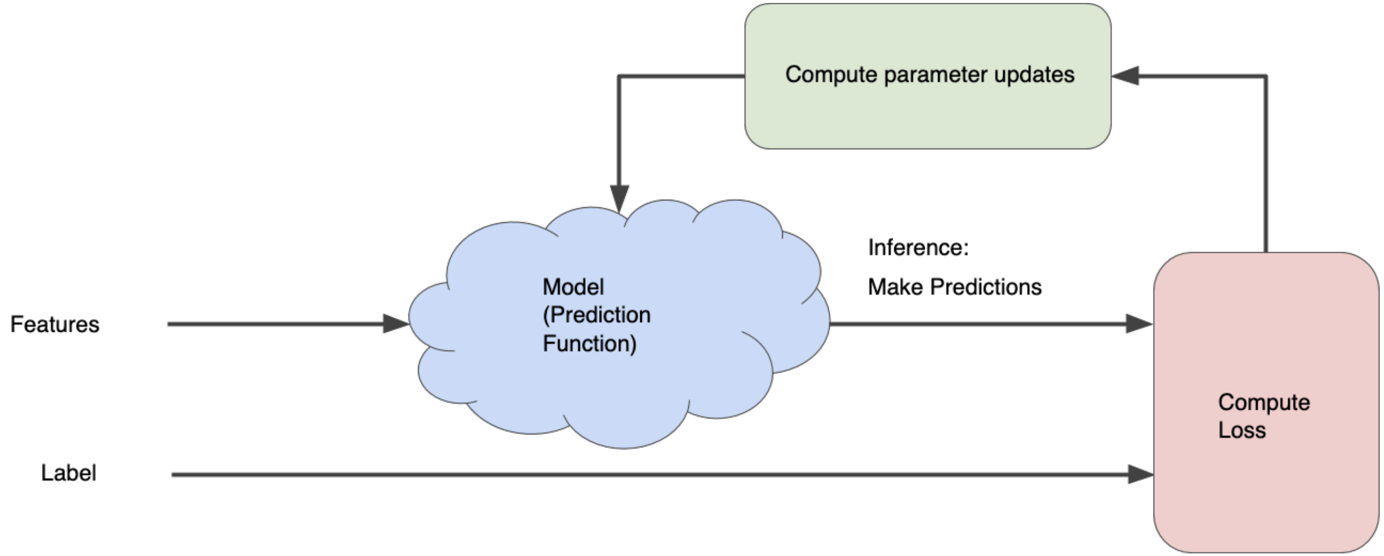
Order or Degree 4

Gradient Descent

- Learning the machine learning model by iteratively reducing the loss

- More like how we learn: *learn by making mistakes*

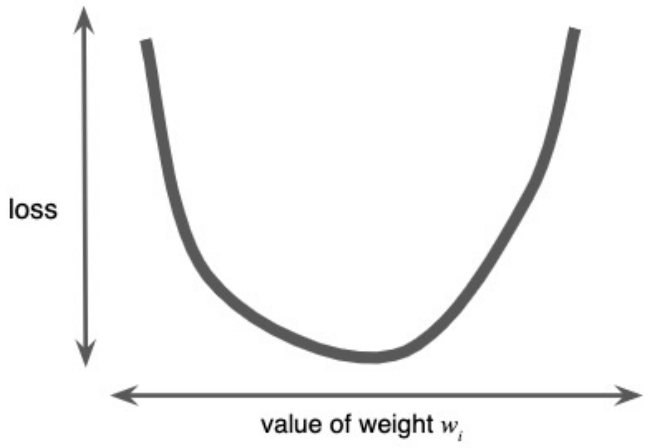
Iteratively Reducing the Loss



https://developers.google.com/machine-learning/crash-course/reducing-loss/an-iterative-approach

Mean Squared Error

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

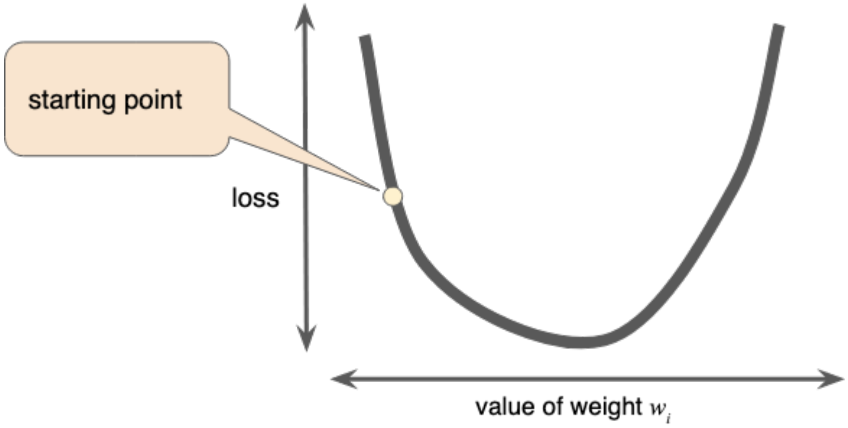


Regression problems yield convex loss vs. weight plots

https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

Gradient Descent (1)

- Pick a starting value of w_i



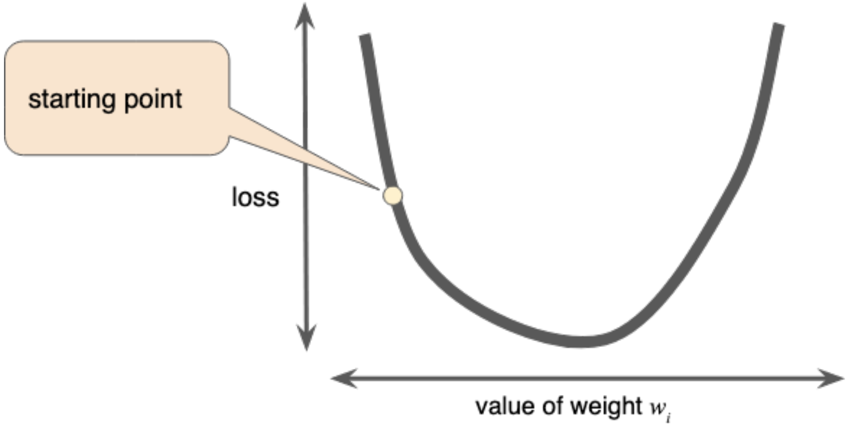
A starting point for gradient descent.

https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

Gradient Descent (2)

- Compute the *gradient* of the loss curve at the starting point

- Gradient of the curve at any point is equal to the *derivative* of the curve at that point



A starting point for gradient descent.

https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

Gradient Descent (3)

- Gradient is a vector: has both magnitude and direction

- It always points to the direction of the steepest increase in the loss function

- What is our objective?
  - Reduce the loss

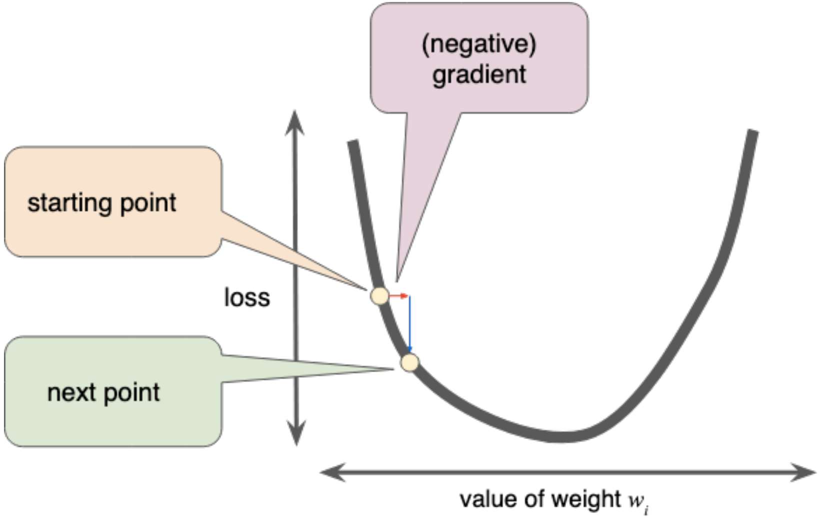
- Thus we can reduce the loss by taking a step in the direction of negative gradient

Gradient Descent (4)

- Reduce the loss by taking a step in the direction of negative gradient

- How much should we move in the desired direction?

    - ➢ some fraction of the gradient's magnitude

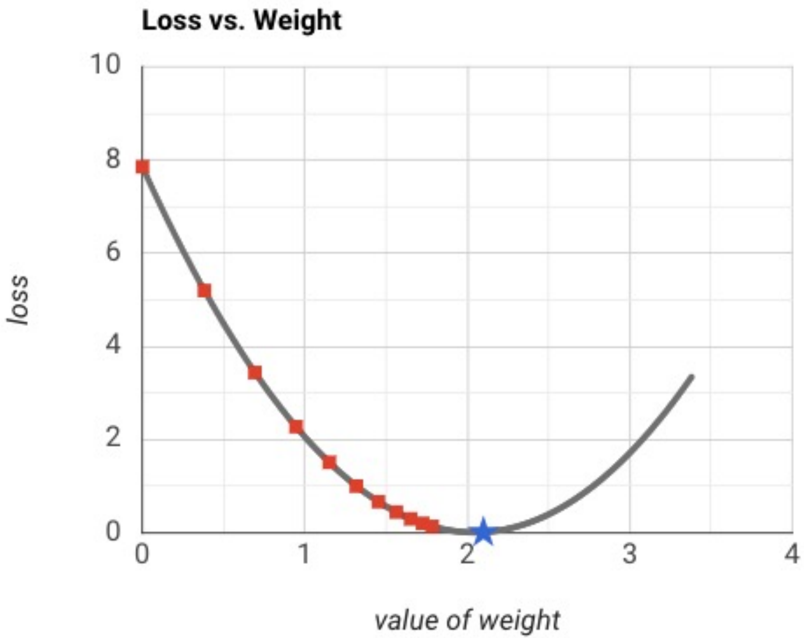


A gradient step moves us to the next point on the loss curve

https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

Gradient Descent (5)

- The gradient descent then repeats this process, edging ever closer to the minimum.



https://developers.google.com/machine-learning/crash-course/fitter/graph

More Formally

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Gradient descent reduces the loss function iteratively

- This is done by updating the values of the parameters as

$$w_0 = w_0 - \alpha\frac{\partial\mathcal{L}}{\partial w_0}$$

$$w_1 = w_1 - \alpha\frac{\partial\mathcal{L}}{\partial w_1}$$

- some fraction of the gradient's magnitude

Known as the *learning rate*

Partial Derivatives of $\mathcal{L}(w_0, w_1)$ (1)

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

- Let's make a slight modification to the loss function just to make the math easier, later on

$$\mathcal{L}(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

Partial Derivatives of $\mathcal{L}(w_0, w_1)$ (2)

$$\mathcal{L}(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

- Let's expand the left hand side

$$\mathcal{L}(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}\left(w_1^2 x_i^2 + 2w_1 x_i w_o - 2w_1 x_i y_i + w_0^2 - 2w_0 y_i + y_i^2\right)$$

$$\frac{\partial \mathcal{L}}{\partial w_0}(1)$$

$$\mathcal{L}(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}\left(w_1^2 x_i^2 + 2w_1 x_i w_o - 2w_1 x_i y_i + w_0^2 - 2w_0 y_i + y_i^2\right)$$

- Let's simplify by removing the terms that do not involve w_0

$$= \frac{1}{2n}\sum_{i=1}^{n}(2w_1 x_i w_o + w_0^2 - 2w_0 y_i)$$

- Now, we are ready to compute the partial derivative

$$\frac{\partial \mathcal{L}}{\partial w_0} = \frac{1}{2n}\sum_{i=1}^{n}(2w_1 x_i + 2w_0 - 2y_i)$$

$$\frac{\partial \mathcal{L}}{\partial w_0} \text{ (2)}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = \frac{1}{2n} \sum_{i=1}^{n} (2w_1 x_i + 2w_0 - 2y_i)$$

- The 2’s cancel out (see, that’s why we included it earlier)

$$\frac{\partial \mathcal{L}}{\partial w_0} = \frac{1}{n} \sum_{i=1}^{n} (w_1 x_i + w_0 - y_i)$$

- We can rewrite it as

$$\frac{\partial \mathcal{L}}{\partial w_0} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

$$\frac{\partial \mathcal{L}}{\partial w_1}$$

$$\mathcal{L} = \frac{1}{2n}\sum_{i=1}^{n}\left(w_1^2 x_i^2 + 2w_1 x_i w_o - 2w_1 x_i y_i + w_0^2 - 2w_0 y_i + y_i^2\right)$$

- Next we will calculate the partial derivative of the loss with respect to w_1
- You will do it yourself: apply the same procedure we used for w_0
- The outcome should be

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\, x_i$$

Thus

$$\mathcal{L}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Gradient descent reduces the loss function iteratively.

- This is done by updating the values of the parameters as

$$w_0 = w_0 - \alpha\frac{\partial\mathcal{L}}{\partial w_0} = w_0 - \alpha\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

$$w_1 = w_1 - \alpha\frac{\partial\mathcal{L}}{\partial w_1} = w_1 - \alpha\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\, x_i$$

GD Summary (Single Predictor)

Repeat {

$$w_0 = w_0 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

$$w_1 = w_1 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)\, x_i$$

}

Multiple Predictors

- In real-life, we most deal with situations where a response variable depends on multiple predictors

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots w_9 x_9$$

- Luckily, this requires a simple update to GD

Recall: GD Summary (Single Predictor)

Repeat {

$$w_0 = w_0 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

$$w_1 = w_1 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)\, x_i$$

}

I can Rewrite it as

Repeat {

$$w_0 = w_0 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)\, x_i^0$$

$$w_1 = w_1 - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)\, x_i^1$$

}

GD Summary (Multiple Predictors)

Repeat {

$$w_j = w_j - \alpha \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)\, x_i^j$$

}

Classification

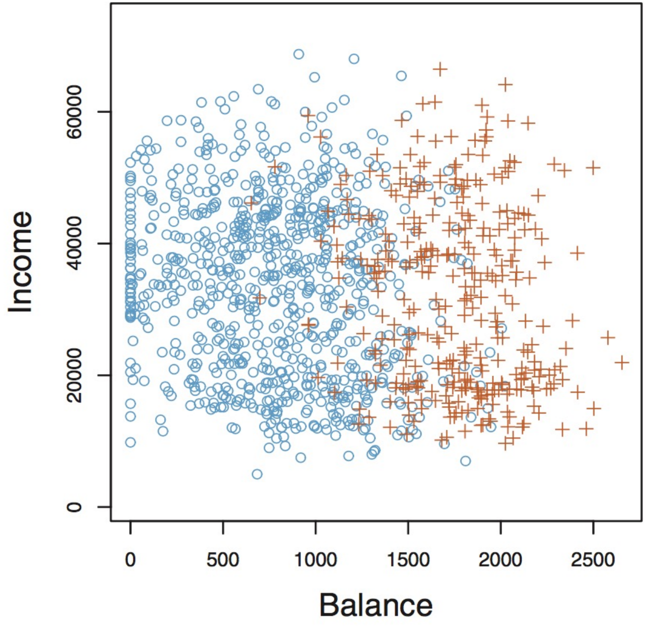
- Response is qualitative, discrete or categorical.

- For examples,
  - Eye color ∈ $\{brown, blue, green\}$
  - Email ∈ $\{Spam, Ham\}$
  - Expression ∈ $\{Happy, Sad, Angry\}$
  - Action ∈ $\{Walking, Running, Cycling\}$

Binary Classification

- Classification task that has two class labels

- For example,
  - People who defaulted on their credit card payments and those who did not



What if we Treat the Problem As Predicting Class Probability?

- Instead of predicting the class of a data point, what if we computed the probability of the data point belonging to a certain class?
- That is $p(Default = Yes \mid balance, income)$
- We can “threshold” the result to decide the class
- What can be the advantage of doing this?
  - y becomes continuous
  - Thus, we can solve this as a regression problem (reuse of what we learned earlier)

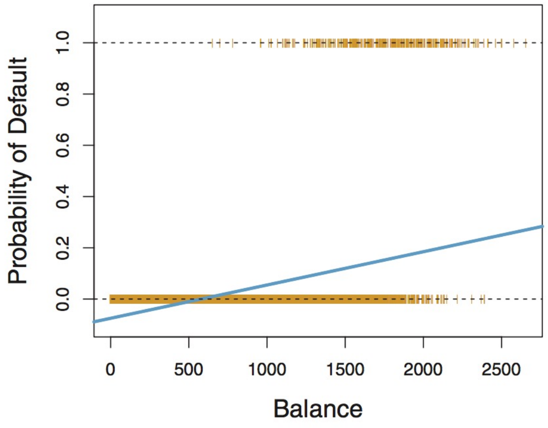
Can we use Linear Regression?

- For simplicity, let's work with a single predictor, and represent "default" with 1

$$p(y = 1 \mid balance)$$

$$= w_o + w_1 \times balance$$

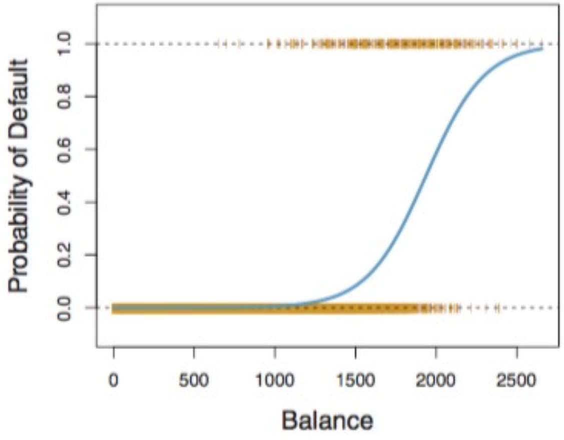
$$p(x) = w_o + w_1 \times balance$$



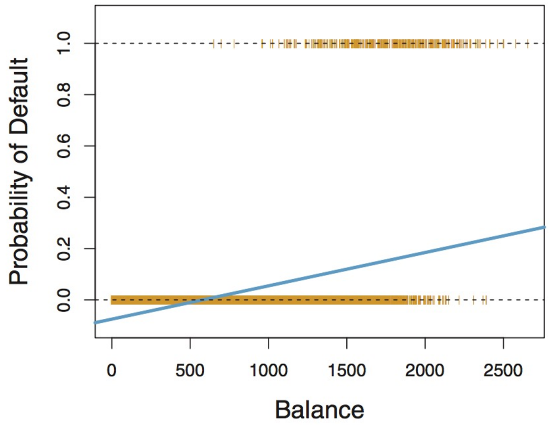
- Do you see any problem?

Limitations of using Linear Regression

- This is what we want



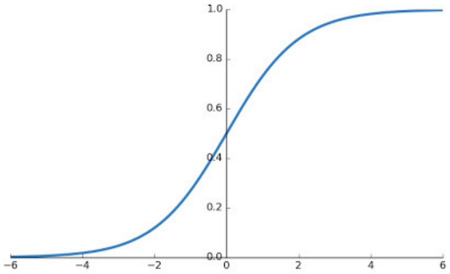
- This is what we get



Logistic Regression

- Thus, we must model $p(x)$ using a function that gives outputs between 0 and 1 for all values of x.

- One such function is the *Logistic or sigmoid function*



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- Thus, in logistic regression, we use the **Logistic Function**

Logistic Regression (Single Predictor)

$$p(x_1) = \frac{1}{1 + e^{-z}}$$

$$z = w_0 + w_1 x_1$$

Logistic Regression (Multiple Predictors)

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-z}}$$

$$z = w_0 + w_1 x_1 + \cdots + w_p x_p$$

Estimating Parameters of Logistic Regression

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-z}}$$

$$z = w_0 + w_1 x_1 + \cdots + w_p x_p$$

We need a *loss function*

A Short Detour

- Linear sum as a dot product

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots w_d x_d$$

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$y = \boldsymbol{w}^T . \boldsymbol{x}$$

Loss Function of Logistic Regression

$$y = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}^i}}$$

- First take a look at linear regression loss function

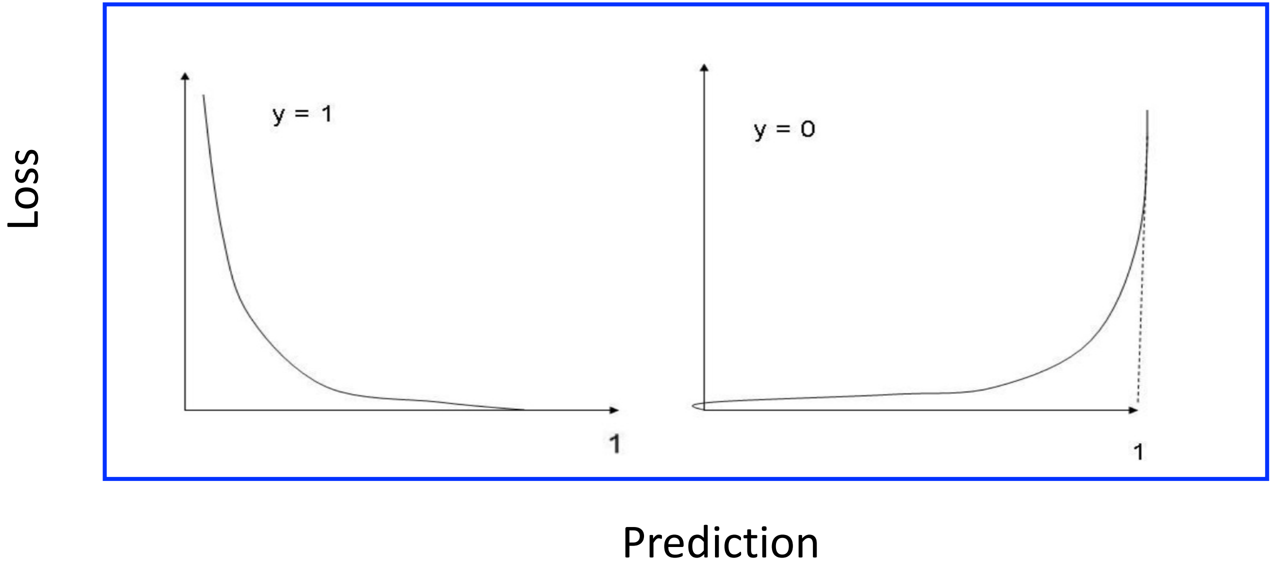
$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(y^i - \boldsymbol{w}^T \boldsymbol{x}^i\right)^2$$

- Due to non-linearity of the sigmoid function in hypothesis of Logistic Regression, MSE is **not convex** anymore

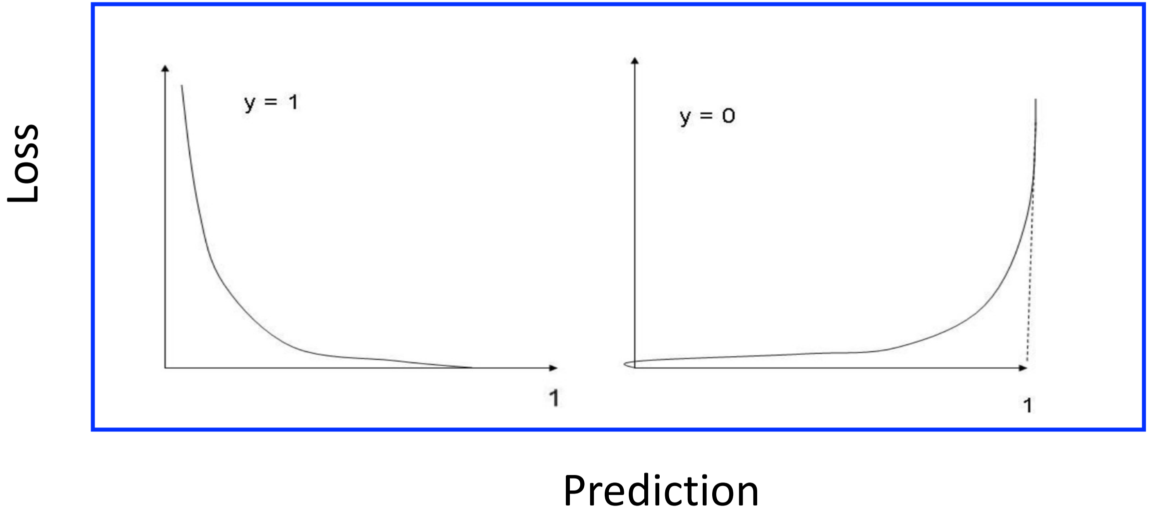
Thus we need a new Loss Function

Loss Function of Logistic Regression (2)

- We want to have soemthing, the behaves like this



Loss Function of Logistic Regression (3)



$$\mathcal{L}(\boldsymbol{w}) = \begin{cases} -\log(p(x)), & if\, y = 1 \\ -log(1 - p(x)), & if\, y = 0 \end{cases}$$

Loss Function of Logistic Regression (4)

$$\mathcal{L}(\boldsymbol{w}) = -\frac{1}{n}\sum_{i=1}^{n} y^i \log\left(p(x^i)\right) + (1 - y^i)\log(1 - p(x^i))$$

- Given this loss function, what do you think we are gonna do next?
- We will define our objective function

$$\boxed{\underset{w_o, w_1}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{w})}$$

Solving Our Objective

- Given our objective function, how would we solve it?

  - ➢ It is done using **Gradient Descent**

- Thus, what we need to do is to derive the equation for parameter update

$$w_j = w_j - \alpha \frac{\partial \mathcal{L}}{\partial w_j}$$

A Short Detour, again!

- We will use the following facts, note $\sigma(.)$ is the logistic or sigmoid function

$$\frac{d\left(\sigma(x)\right)}{dx} = \sigma(x)\left(1 - \sigma(x)\right)$$

$$\frac{d\left(log(x)\right)}{dx} = \frac{1}{x}\frac{d}{dx}x$$

$$\frac{d\left(f\left(g(x)\right)\right)}{dx} = f'\left(g(x)\right)g'(x)$$

Deriving the Equation for Parameter Update

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log\big(p(x_i)\big) + (1 - y_i)\log(1 - p(x_i))$$

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log\big(\sigma(z_i)\big) + (1 - y_i)\log(1 - \sigma(z_i))$$

$$z_i = \boldsymbol{w}^T \boldsymbol{x}^i = w_0 + w_1 x_1^i + \cdots + w_p x_p^i$$

Deriving the Equation for Parameter Update (2)

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \boxed{\log(\sigma(z_i))} + (1 - y_i)\boxed{\log(1 - \sigma(z_i))}$$

- Computing $\frac{\partial \mathcal{L}}{\partial w_j}$
- These are what depend on w_j
- So let's compute their partial derivative w.r.t w_j

Deriving the Equation for Parameter Update (3)

$$\frac{\partial \log(\sigma(z_i))}{\partial w_j} = \frac{1}{\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial z_i} \frac{\partial z_i}{\partial w_j}$$

- Which rules did we apply?
  - Rule of log
  - Chain rule

Deriving the Equation for Parameter Update (4)

$$\frac{\partial \log\big(\sigma(z_i)\big)}{\partial w_j} = \frac{1}{\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial z_i} \frac{\partial z_i}{\partial w_j}$$

$$= \frac{1}{\cancel{\sigma(z_i)}} \cancel{\sigma(z_i)}\big(1 - \sigma(z_i)\big) \frac{\partial z_i}{\partial w_j}$$

- Which rule did we apply?
  - Rule of sigmoid function

Deriving the Equation for Parameter Update (5)

$$\frac{\partial \log\big(\sigma(z_i)\big)}{\partial w_j} = \big(1 - \sigma(z_i)\big)\frac{\partial z_i}{\partial w_j}$$

$$z_i = \boldsymbol{w}^T \boldsymbol{x}^i = w_0 + w_1 x_1^i + \cdots + w_p x_p^i$$

$$\frac{\partial z_i}{\partial w_j} = x_j^i$$

$$\boxed{\frac{\partial \log\big(\sigma(z_i)\big)}{\partial w_j} = \big(1 - \sigma(z_i)\big)x_j^i}$$

Recall

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \boxed{\log\left(\sigma(z_i)\right)} + (1 - y_i)\boxed{\log(1 - \sigma(z_i))}$$

- Computing $\frac{\partial \mathcal{L}}{\partial w_j}$
- These are what depend on w_j
- So let's compute their partial derivative w.r.t w_j

Deriving the Equation for Parameter Update (6)

$$\frac{\partial \log(1 - \sigma(z^i))}{\partial w_j} = ?$$

- You will do this activity yourselves
- You just have to apply the same rules

Deriving the Equation for Parameter Update (7)

$$\frac{\partial \log(1-\sigma(z^i))}{\partial w_j} = -\sigma(z^i)x_j^i$$

Deriving the Equation for Parameter Update (8)

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \boxed{\log(\sigma(z_i))} + (1-y_i)\boxed{\log(1-\sigma(z_i))}$$

$$\frac{\partial \log(\sigma(z_i))}{\partial w_j} = (1-\sigma(z_i))x_j^i$$

$$\frac{\partial \log(1-\sigma(z^i))}{\partial w_j} = -\sigma(z^i)x_j^i$$

Deriving the Equation for Parameter Update (8)

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i \boxed{\log(\sigma(z_i))} + (1 - y_i)\boxed{\log(1 - \sigma(z_i))}$$

- Thus

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n} y^i\left(1 - \sigma(z^i)\right) x_j^i - (1 - y^i)\sigma(z^i)x_j^i$$

Deriving the Equation for Parameter Update (9)

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n} y^i \left(1 - \sigma\left(z^i\right)\right) x_j^i - \left(1 - y^i\right)\sigma\left(z^i\right)x_j^i$$

- Solving it further by multiplying and simplifying, we will get

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n} \left(y^i - \sigma\left(z^i\right)\right) x_j^i$$

Deriving the Equation for Parameter Update (10)

$$\frac{\partial \mathcal{L}}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n}\left(y^i - \sigma(z^i)\right) x_j^i$$

- We can rewrite it as

$$\frac{\partial \mathcal{L}}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}\left(p(x^i) - y^i\right) x_j^i$$

Deriving the Equation for Parameter Update (11)

- Thus,

$$w_j = w_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(p(x^i) - y^i \right) x_j^i$$

Note

- We move over all data, average the result and then update the parameter

$$w_j = w_j - \alpha \boxed{\frac{1}{n}\sum_{i=1}^{n}} \left(p(x^i) - y^i\right) x_j^i$$

Types of Gradient Descent

- Batch GD
- Stochastic GD
- Mini-batch GD

From Probability to Class Label

$$\hat{y} = \begin{cases} 1, & \hat{p}(\boldsymbol{x}) > Th \\ 0, & otherwise \end{cases}$$

Metrics (1)

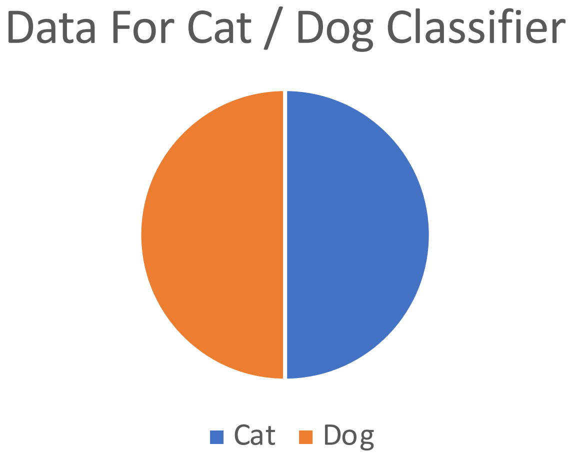
- Accuracy

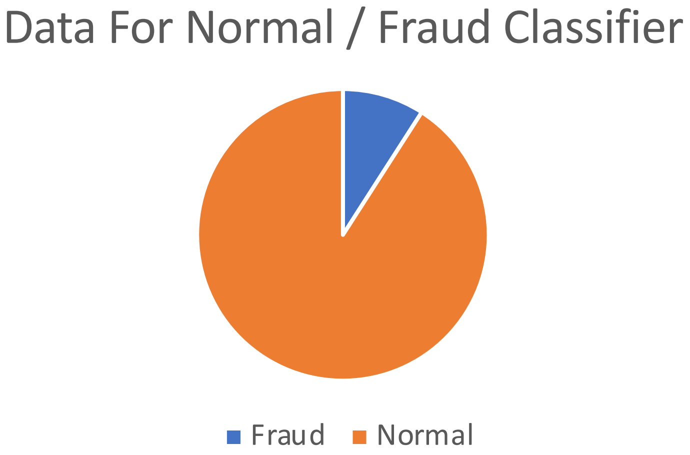
$$Acc = 1 - \frac{\#\ of\ missclassified\ examples}{\#\ of\ samples}$$

- But sometimes, accuracy is not a good metric to use to measure performance of machine learning models

Things to Know about Machine Learning

- Class-imbalance





Example

Confusion Matrix

1. How accurate is this classifier?

Predicted Class Label

Actual Class Label		**Negative**	**Positive**
	Negative	998	0
	Positive	2	0

2. But what if the missclassified data points (False Negatives) are someone with a serious disease, or a serious fraudalent transcation, or a bad malware, etc.

General Requirements for ML Models

1. Better than random guessing

2. Better than majority guessing

Confusion Matrix

		Predicted Class Label	
		Negative	**Positive**
Actual Class Label	Negative	True Negatives	False Positives
	Positive	False Negatives	True Positives

Metrics (2)

		Predicted Class Label	
		Negative	Positive
Actual Class Label	Negative	True Negatives	False Positives
	Positive	False Negatives	True Positives

1. Precision

$$\color{purple}{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Recall

$$\color{orange}{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Back to Our Example

		Predicted Class Label	
		Negative	Positive
Actual Class Label	Negative	998	0
	Positive	2	0

1. Precision = 0

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Recall = 0

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Our Example–Case 2

		Predicted Class Label	
		Negative	Positive
Actual Class Label	Negative	998	0
	Positive	0	2

1. Precision = 1

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Recall = 1

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Our Example–Case 3

		Predicted Class Label	
		Negative	Positive
Actual Class Label	Negative	998	0
	Positive	1	1

1. Precision = 1

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Recall = 0.5

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Balancing Precision and Recall

- F1 Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Summary

- Gradient Descent
- Classification Tasks
- Logistic Regression and its learning objective
- Solving the objective
  - Driving the equation for parameter updates
  - Gradient Descent
- Classification metrics and class imbalance