1. Explain what logistic regression is.

logistic regression یکی از مهم ترین روش های classification در یادگیری ماشین می باشد. برخلاف linear regression که خروجی می توانست مقادیری بیوسته داشته باشد، در logistic regression خروجی به یکی از چند مقدار محدود (class/category) تعلق خواهد داشت. به عنوان مثال در یک مسئله Binary classification ما دو مقدار 1 و 0 (True/false, Positive/negative) را خواهیم داشت.

sigmoid function / logistic function بصورت زیر تعریف می شود:

$$f_{\vec{w},b}(\vec{x}) = g(z) = \frac{1}{1+e^{-z}} \quad ; \quad z = \vec{w}\vec{x} + b$$

خروجی تابع فوق همواره عددی بین 0 و 1 خواهد بود. در واقع این عدد احتمال تراکیبی داده ورودی (مجموعه ویژگی های آن یعنی بردار x) در کلاس 1 را نشان می دهد. برای این منظور treshhold ای تعیین می شود. (فرضا 0.5) که اگر احتمال مورد نظر از آن treshhold بیشتر شد، داده ورودی به کلاس 1 / Positive / True تعلق دارد.

$$f_{\vec{w},b}(\vec{x}) = P(y=1 \mid \vec{x}; \vec{w}, b)$$

به عنوان مثال در سوال مطرح شده که شامل تعدادی نقاط با رنگ های متفاوت است، ما به دنبال معنی هستیم که با کمترین مقدار خطا مرز بین نقاط با رنگ مختلف را مشخص کند. در حالت اول مثال ما به دنبال یک خط به فرم w_1x+b بودیم که چون یک خط نمی توانست این دسته نقاط را از هم جدا کند، دایره ای با درجه بالاتر از 1 را انتخاب کردیم که دقت Accuracy بالایی بدست می آورد. همواره بدست آوردن معادله z و قرار دادن آن در معادله sigmoid هرچه نقاط از معادله z دورتر باشند، احتمال بالاتری برای تفکیک در کلاس وجود دارد. رابطه خطا هم بصورت زیر می باشد:

$$J(\vec{w},b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) + (1-y^{(i)})\log\left(1-f_{\vec{w},b}(\vec{x}^{(i)})\right)\right]$$

2. write its formula for this Problem --- 9 3.

$$z = w_1x_1 + w_2x_2 + b = 0 \xrightarrow[\text{boundry}]{\text{decision}} x_2 = -1/1x_1 + 20$$

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

2. Explain the following metrics:

accuracy: $\frac{TP+TN}{P+N} = \frac{\text{No. Correct Predictions}}{\text{No. All Predictions}}$

از میان کل پیش بینی ها چه تعداد درست پیش بینی شده اند. یعنی مجموع تعداد داده هایی که به درستی به کلاس ۰ و ۱ تخصیص داده شده اند تقسیم بر کل داده هایی که برای آن ها Label تعیین کرده است.

Precision: $\frac{TP}{TP+FP}$ از میان کل پیش بینی های Positive، یعنی تمامی داده هایی که به کلاس هدف ما تخصیص داده شده اند، چه تعداد به درستی پیش بینی شده اند.

Recall: $\frac{TP}{TP+FN}$ از میان کل داده هایی که واقعاً Positive Label داشته، چه تعدادی را مدل ما پیدا کرده است.

Specificity: True Negative Rate

$\frac{TN}{TN+FP}$ همان Recall اما Label هدف، Negative است. از میان کل داده های Negative چه تعدادی به درستی پیش بینی شده اند.

Sensitivity: True Positive Rate = Recall

* متادیر Metric ها در فایل Colab *

3. Problems that Recall is more important than Precision?

طبعاً در مسائلی که پیدا کردن همه متادیر Positive برای ما بسیار حیاتی است و هزینه FN بسیار بیشتر از FP است.

مثال: تشخیص این که یک تومور سرطانی است.

4. Problems that Precision is more important than recall?

در مسائلی که FP بسیار پرهزینه تر از FN می باشد.

مثال: تشخیص ایمیل Spam.

Problem 2)

1. write the loss function for this formula.

$$J(\vec{w},b) = \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2$$

→ This is cost function, loss is simply this formula but for one sample, so we don't have $\frac{1}{2m}$ & Σ.

for this Problem we have

$$J(w_2,w_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(w_2 x_2^{(i)^2} + w_1 x_1^{(i)} - y^{(i)}\right)^2$$

2. Drive the derivation of the loss function.

Gradient descent

repeat {

$$\frac{dJ}{dw_1} = \frac{1}{m}\sum_{i=1}^{m}\left(w_2 x_2^{(i)^2} + w_1 x_1^{(i)} - y^{(i)}\right) x_1^{(i)}$$

$$\frac{dJ}{dw_2} = \frac{1}{m}\sum_{i=1}^{m}\left(w_2 x_2^{(i)^2} + w_1 x_1^{(i)} - y^{(i)}\right) x_2^{(i)^2}$$

}

3. Explain the impact of input normalization ---

وقتی ورودی (input) نرمال سازی نشده باشد: مثلاً feature اول در بازه [0,500] و feature دوم در بازه [0,1] باشد، در این صورت روش gradient descent در مسیر رسیدن به بهینه نوسان (oscillate) می کند. در نتیجه دیر تر Converge می کند. علت این موضوع این است که تغییرات در راستای یک feature بسیار بزرگتر از feature دیگر خواهد بود، همان مشتق بر اساس یک feature

Problem 3)

Step 1.

1. Explain why does overfitting happen?

یکی از دلایلی که می تواند منجر به overfitting شود: ۱) کم بودن تعداد Sample های Dataset.

۲) بالا بودن تعداد feature ها.

۳) بالا بودن پیچیدگی مدل و واریانس بالا.

در این مثال به علت درجه ۲۰ بودن، پیچیدگی مدل بالا بوده و مدل دارای واریانس بالا است. به این معنی که با کوچکترین تغییر در یک داده مدل ما به شکل قابل توجهی تغییر می کند. همچنین تعداد داده های بسیار کم است. در نهایت مدل می تواند generalize نشود و بسیار وابسته به داده های train ی شود.

2. Discuss about the training error and gereralization error of this model.

training error همان error روی training Dataset ما می باشد. از آنجایی که تعداد Data های ما ۲۱ بوده و مدل ما از مرتبه ۲۰ است. کلیه داده ها را دقیق استاده و عملاً training error ما حتی خود صفر است. اما gereralization error میزان دقت مدل روی داده های از قبل دیده نشده را بررسی می کند. این تعداد می تواند با جلوگیری از overfitting افزایش یابد. طبیعی overfitting باعث می شود تا پیش بینی مدل کاهش پیدا کند و به اصطلاح مدل generalize نباشد. بنابراین gereralization error افزایش می یابد.

Answer these questions:

1. همانطور که مشاهده می شود. با استفاده از regularization تا میزان خوبی از overfitting جلوگیری کردیم و به ست generalize بودن مدل پیش رفتیم. در واقع regularization با وارد کردن پارامتر های مدل به cost function سعی در کاهش وزن ها می کند که باعث می شود تا مدل overfit نشود. همانطور که مشاهده می شود L1 regularization بهترین عملکرد را داشته و از همه بیشتر به تابع اصلی x^3 نزدیک است.

2.

$$J(\vec{w},b) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2 + \begin{cases} \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2 \rightarrow \text{Ridge} \\ \frac{\lambda}{2m}\sum_{j=1}^{n} w_j \rightarrow \text{Lasso} \end{cases}\right]$$

Lasso ← برای Feature selection استفاده می‌شود و ضریب برخی از Feature ها را صفر می‌کند.

Ridge ← سعی می‌کند تاثیر این Feature ها را به شکل یکنواخت‌تری کم کند.

در واقع تفاوت Ridge و Lasso در Penalized term آن‌ها می‌باشد.

3. همانطور که گفته شد L1(Lasso) در این مثال بهتر عمل کرده است. زیرا Lasso برخی از Feature ها را صفر کرده که باعث می‌شود تا با قدرت کمتری Feature ، generalization بهتری داشته باشیم.

4. در histogram ها مشاهده می‌شود که در L2 توزیع یکنواخت‌تری نسبت به histogram بدون Regularization داریم، درصورتی که در L1 باعث Feature selection برخی Feature ها حذف شده و مقادیر آن‌ها بیشتر در صفر است که باعث می‌شود مدل از مرتبه ۲ کم اثرتر شود.

5. بی‌شک L1 ، در واقع این موضوع مهم‌ترین برتری L1 regularization است، چون باعث ایجاد Sparsity در ضرایب می‌شود و Feature های کم اهمیت‌تر و noise ها را حذف می‌کند. در نتیجه در برابر Outlier ها قدرتمندتر است.