# 기초과제 2 모범답안

머신러닝을 위한 수학 + 전처리 기초

작성자 : 10기 신재우

## 1 전처리 기초

### 1.1 .ipynb 파일을 참조해서 state.name 컬럼에서 결측값들을 모두 삭제해 주시오.

```
turnout = turnout.dropna(subset = 'state.name')
na_state = turnout['state.name'].isna().any()
print("NaN State :", na_state)
```

```
NaN State :  False
```

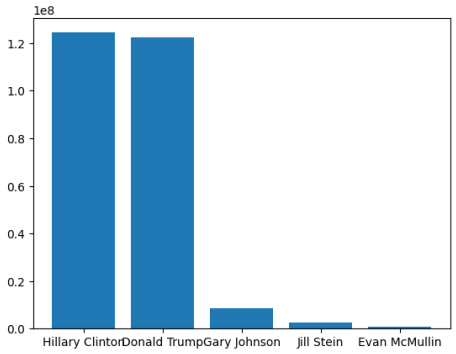
### 1.2 .ipynb 파일을 참조해서 가장 많은 투표 수를 얻은 5명의 후보자들을 구하시오.

```
top_5 = turnout.groupby('cand')['votes'].sum().sort_values(ascending = False).head(5)
print(top_5)
```

```
cand
Hillary Clinton    124430025
Donald Trump       122236995
Gary Johnson         8526205
Jill Stein           2619645
Evan McMullin         939540
Name: votes, dtype: int64
```

### 1.3 .ipynb 파일을 참조해서 가장 많은 투표 수를 얻은 5명의 후보자들에 대한 데이터를 시각화 해주시오.

```
plt.bar(top_5.index, top_5)
plt.show()
```




# 2 최적화개론

## 2.1 Logistic Regression 의 Convex Optimization 과정을 서술하시오.

Logistic Regression 의 Objective Function 식을 다음과 같이 쓰겠습니다.

$$\min_w J(w) = \min_w \sum_{i=1}^{m} \left\{ -y^{(i)} \log\left(\hat{y}^{(i)}\right) - \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) \right\}$$

이것을 서술하기 위해서는 이 식 자체가 Convex 하다는 것을 먼저 증명해주고 난 다음에 Convex 한 Problem (문제) 를 푸는 방법은 Gradient Descent Method 등의 최적화 방법을 통해서 풀 수가 있습니다. 우선은 $J(w)$ 는 다음과 같이 두 식으로 나눠줄 수가 있습니다.

$$-y^{(i)} \log\left(\hat{y}^{(i)}\right), \quad -\left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)$$

식의 $\hat{y}^{(i)}$ 은 Logistic Function 식을 활용하면 다음과 같이 표현이 가능합니다.

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

이것을 $J(w)$ 식에 적용해서 좌측과 우측 식을 다음과 같이 나눠줄 수 있습니다.

$$-y^{(i)} \log\left(\frac{1}{1 + e^{-w^T x^{(i)}}}\right), \quad -\left(1 - y^{(i)}\right) \log\left(1 - \frac{1}{1 + e^{-w^T x^{(i)}}}\right)$$

이 두 식들이 Convex 하다는 것을 증명하면 $J(w)$ 역시 Convex 하다는 것을 알 수가 있습니다.

우선은 좌측 식을 살피겠습니다.

$$-y^{(i)} \log \left( \frac{1}{1+e^{-w^T x^{(i)}}} \right)$$

저희는 $w$ 에 대해서 Convex 한 것을 증명해야 합니다. 이것은 $w$ 에 대해서 2 번 미분을 시키면 되며, 이것은 다음과 같습니다.

$$-\frac{\partial}{\partial w} y^{(i)} \log \left( \frac{1}{1+e^{-w^T x^{(i)}}} \right) = -y^{(i)} \cdot \frac{1}{\frac{1}{1+e^{-w^T x^{(i)}}}} \cdot \frac{\partial}{\partial w} \left[ \frac{1}{1+e^{-w^T x^{(i)}}} \right]$$

$$= -y^{(i)} \cdot \left(1+e^{-w^T x^{(i)}}\right) \cdot \frac{-1}{\left(1+e^{-w^T x^{(i)}}\right)^2} \cdot \frac{\partial}{\partial w} \left[1+e^{-w^T x^{(i)}}\right]$$

$$= -y^{(i)} \cdot \frac{-1}{1+e^{-w^T x^{(i)}}} \cdot \left(-x^{(i)} \cdot e^{-w^T x^{(i)}}\right) = -y^{(i)} \cdot x^{(i)} \cdot \frac{e^{-w^T x^{(i)}}}{1+e^{-w^T x^{(i)}}}$$

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

이것에 대해서 한번 더 미분을 하게 된다면 다음과 같이 나오게 됩니다.

$$\frac{\partial}{\partial w} \left[ -y^{(i)} \cdot x^{(i)} \cdot \frac{1}{e^{w^T x^{(i)}}+1} \right] = -x^{(i)} y^{(i)} \cdot \frac{-1}{\left(e^{w^T x^{(i)}}+1\right)^2} \cdot \frac{\partial}{\partial w} \left[ e^{w^T x^{(i)}}+1 \right]$$

$$= x^{(i)} y^{(i)} \cdot \frac{x^{(i)} \cdot e^{w^T x^{(i)}}}{\left(e^{w^T x^{(i)}}+1\right)^2} = \frac{\left(x^{(i)}\right)^2 \cdot y^{(i)} \cdot e^{w^T x^{(i)}}}{\left(e^{w^T x^{(i)}}+1\right)^2}$$

마지막 식을 살펴보면 $x^{(i)}$ 과 분모는 는 제곱으로 항상 0 보다 크거나 같습니다. $y^{(i)}$ 는 0 과 1 의 값들이 있기 때문에 0 보다 크거나 갖습니다. $e^{w^T x^{(i)}}$ 의 경우 $e^x$ 에 관한 함수가 $x$ 에 대해서 항상 0 보다 크거나 같습니다.

결론적으로는 다음과 같은 식을 만족하기 때문에 좌측 식이 Convex 하다는 것을 알 수가 있습니다.

$$\frac{\left(x^{(i)}\right)^2 \cdot y^{(i)} \cdot e^{w^T x^{(i)}}}{\left(e^{w^T x^{(i)}}+1\right)^2} \geq 0$$

이제는 우측 식이 Convex 하다는 것을 증명하겠습니다. 우측 식을 가져오면 다음과 같이 바꿀 수가 있습니다.

$$-\left(1-y^{(i)}\right) \cdot \log \left(1-\frac{1}{1+e^{-w^T x^{(i)}}}\right) = -\left(1-y^{(i)}\right) \cdot \log \left(\frac{1+e^{-w^T x^{(i)}}-1}{1+e^{-w^T x^{(i)}}}\right)$$

$$= -\left(1-y^{(i)}\right) \cdot \log \left(\frac{e^{-w^T x^{(i)}}}{1+e^{-w^T x^{(i)}}}\right) = -\left(1-y^{(i)}\right) \cdot \left[\log \left(e^{-w^T x^{(i)}}\right) - \log \left(1+e^{-w^T x^{(i)}}\right)\right]$$

$$= -\left(1-y^{(i)}\right)\cdot\left[-w^T x^{(i)} + \log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right)\right] = \left(1-y^{(i)}\right)\cdot\left[w^T x^{(i)} - \log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right)\right]$$

$$= \left(1-y^{(i)}\right)\cdot\left(w^T x^{(i)}\right) - \left(1-y^{(i)}\right)\cdot\log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right)$$

이것 역시 2개의 식으로 나눠지는 모습을 볼 수가 있습니다. 좌측 식의 경우 이미 $w$ 에 대해서 Linear 함수, 혹은 다른 말로 Affine Function, 이기 때문에 Convexity 를 이미 만족합니다. 우측 식의 경우 위의 식이었던

$$-y^{(i)}\log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right)$$

식과 다를게 없으며 이것이 이미 Convex 하다는 것을 증명했기 때문에 전체 식인 $J(w)$ 이 Convex 하다는 것을 만족하게 됩니다.

Logistic Regression 문제를 다시 가져오면 다음과 같습니다.

$$\min_w J(w) = \min_w \sum_{i=1}^{m}\left\{-y^{(i)}\log\left(\hat{y}^{(i)}\right) - \left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\right\}$$

Convex 하다는 것을 증명했기 때문에 이것은 Gradient Descent Algorithm 등의 방법들을 이용해서 푸는 것이 가능해집니다.

### 2.2 $t = 2$ 시점 때의 $x_2$ 값을 구하시오

$$f(x) = x^4 - 2x^3 - 3x^2 + x$$

$$\nabla_x f(x) = 4x^3 - 6x^2 - 6x + 1$$

다음과 같은 하이퍼 파라미터들이 주어졌습니다 $x_1 = 1.7, \eta = 0.1$. 이것을 활용해서 $x_2$ 를 구하겠습니다.

$$x_2 = x_1 - \eta\cdot\nabla_x f(x_1)$$

$$= 1.7 - 0.1\cdot\left(4x_1^3 - 6x_1^2 - 6x_1 + 1\right)$$

$$= 1.7 - 0.1\cdot\left(4\cdot 1.7^3 - 6\cdot 1.7^2 - 6\cdot 1.7 + 1\right) = 1.7 - 0.1\cdot(19.652 - 17.34 - 10.2 + 1)$$

$$= 1.7 - 0.1\cdot(-6.888) = 1.7 + 0.6888 = 2.3888$$

## 3 정보이론 기초

### 3.1 spam.csv데이터 를 .ipynb 파일에서 관찰하고 데이터의 Entropy 를 구하시오.

Number of Spams : 8

Number of Not Spams : 12

$$P(X = \text{Spam}) = \frac{8}{20} = 0.4, \quad P(X = \text{Not Spam}) = \frac{12}{20} = 0.6$$

Entropy 의 공식은 다음과 같습니다.

$$H(x) = -\sum_{x \in X} P(x) \log_2 P(x)$$

이것을 $X = \text{Spam}$ 과 $X = \text{Not Spam}$ 으로 나누면 다음과 같이 Entropy 를 계산할 수가 있습니다.

$$H(x) = -P(X = \text{Spam}) \cdot \log_2 P(X = \text{Spam}) - P(X = \text{Not Spam}) \cdot \log_2 P(X = \text{Not Spam})$$

$$= -0.4 \cdot \log_2(0.4) - 0.6 \cdot \log_2(0.6) = -0.4 \cdot -1.322 - 0.6 \cdot -0.737 = 0.4 \cdot 1.322 + 0.6 \cdot 0.737$$

$$= 0.5288 + 0.4422 = 0.971$$

### 3.2 P 를 spam.csv 으로, Q 를 spam_other.csv 으로 두며 두 분포의 차이를 KL-Divergence 를 이용해서 차이를 손으로 계산해주세요.

정리를 하자면 P 와 Q 들의 값을 다음과 같습니다.

$$P(X = \text{Spam}) = 0.4, \quad P(X = \text{Not Spam}) = 0.6$$

$$Q(X = \text{Spam}) = 0.6, \quad Q(X = \text{Not Spam}) = 0.4$$

KL-Divergence 의 식은 다음과 같습니다.

$$KL(P \| Q) = \sum_{x \in X} P(x) \log \left( \frac{P(x)}{Q(x)} \right)$$

이것을 적용시키면 다음과 같이 나오게 됩니다.

$$KL(P \| Q) = P(X = \text{Spam}) \cdot \log \left( \frac{P(X = \text{Spam}}{Q(X = \text{Spam})} \right) + P(X = \text{Not Spam}) \cdot \log \left( \frac{P(X = \text{Not Spam})}{Q(X = \text{Not Spam})} \right)$$

$$= 0.4 \cdot \log \left( \frac{0.4}{0.6} \right) + 0.6 \cdot \log \left( \frac{0.6}{0.4} \right) = 0.4 \cdot \log(2/3) + 0.6 \cdot \log(3/2)$$

$$= 0.4 \cdot -0.4054651 + 0.6 \cdot 0.4054651 = 0.081093$$

### 3.3 KL-Divergence 은 Asymmetric 한 특징을 가지고 있습니다. 이것을 증명해주시오.

만약에 Symmetric 하다고 가정을 지어보겠습니다. 즉 다음과 같은 공식이 해당이 된다는 가정입니다.

$$KL(P \| Q) = KL(Q \| P)$$

이것의 식을 풀어보면 다음과 같습니다.

$$KL(P \| Q) = \sum_{x \in X} P(x) \log \left( \frac{P(x)}{Q(x)} \right)$$

$$KL(Q\|P) = \sum_{x\in X} Q(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

$$\sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right) = -\sum_{x\in X} P(x)\log\left(\frac{Q(x)}{P(x)}\right) \neq \sum_{x\in X} Q(x)\log\left(\frac{Q(x)}{P(x)}\right)$$

그렇기 때문에 다음과 같이 KL-Divergence 는 Asymmetric 하게 됩니다.

$$KL(P\|Q) \neq KL(Q\|P)$$

# 4 수리통계학

## 4.1 베이즈 정리를 활용해서 DSL 학회원이 아닌 교내생이 감기에 걸렸을 확률을 구하시오.

다음과 같이 $A$ 와 $B$ 를 지정해서 풀겠습니다.

$$A : \text{감기}, \quad B : \text{DSL 학회원}$$

$$P(A) = 0.1, \quad P(B) = 0.001, \quad P(A \mid B) = 0.01$$

저희가 궁금한 것은 DSL 이 아닌 교내생이 감기에 걸렸을 확률입니다. 여기에서 DSL 이 아닌 것이 이미 주어졌습니다. 즉 저희는 다음과 같은 식을 구하는 것이 목표입니다.

$$P(A \mid B^c)$$

이것은 다음과 같이 우선은 $P(B \mid A)$ 를 구해야 합니다.

$$P(B \mid A) = \frac{P(A \mid B)\cdot P(B)}{P(A)} = \frac{0.01 \cdot 0.001}{0.1} = 0.0001$$

다음으로는 $P(B^c \mid A)$ 를 구해야 하며, 저희는 $P(B \mid A)$ 에 관한 값이 이미 있기 때문에 둘의 합은 1이기에 이 점을 활용해서 구하면 됩니다.

$$P(B^c \mid A) = 1 - P(B \mid A) = 1 - 0.0001 = 0.9999$$

해당 식은 다음과 같이 표현이 가능합니다.

$$P(B^c \mid A) = \frac{P(B^c \cap A)}{P(A)} = \frac{P(A \cap B^c)}{0.1} = 0.9999$$

$$P(A \cap B^c) = 0.9999 \cdot 0.1 = 0.09999$$

다음으로는 $P(B^c)$ 값을 구해야 하며, $P(B)$ 에 대한 값이 있기 때문에 이를 활용해서 풀면 됩니다.

$$P(B^c) = 1 - P(B) = 0.999$$

마지막으로는 $P(A \mid B^c)$ 를 구하면 되며, 이것은 다음과 같이 구할 수가 있습니다.

$$P(A \mid B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{0.09999}{0.999} = 0.1000900901$$

번외로 이 문제를 이렇게 어렵게 낼 계획은 없었는데 이렇게 됐다는 점 죄송합니다...

## 4.2 다음과 같은 문제에서 $\theta$ 의 MLE 를 구하시오.

$X_1, \cdots, X_n$ 은 i.i.d 하며 다음과 같은 pdf 를 가집니다.

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad 0 < \theta < \infty$$

이때 $\theta$ 의 MLE 인 $\hat{\theta}$ 를 구하시오.

우선은 Regulatory Conditions 들을 모두 만족하기 때문에 우도함수를 구하면서 미분하는 과정으로 MLE 를 구하겠습니다.

처음엔 우도함수를 만들어야하며, 이는 다음과 같이 만들 수가 있습니다.

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1}$$

로그를 씌우게 되면 다음과 같이 나오게 됩니다.

$$l(\theta) = \sum_{i=1}^{n} \log\left(\theta x_i^{\theta-1}\right) = \sum_{i=1}^{n} \left(\log(\theta) + (\theta-1)\log(x_i)\right)$$

이것을 $\theta$ 에 대해서 미분하게 된다면 다음과 같이 나오게 됩니다.

$$\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \left(\frac{1}{\theta} + \log(x_i)\right) = 0$$

$$\sum_{i=1}^{n} \frac{1}{\theta} = \frac{n}{\theta} = -\sum_{i=1}^{n} \log(x_i)$$

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(x_i)}$$

위와 같이 MLE 를 구할 수가 있게 됩니다. 엄밀히 따지자면 한번 더 미분시켜서 Convex 한지, 즉 이것이 Maximum 한지에 대해서도 구해야 하지만 넘기겠습니다.

과제 하시느라 너무 고생하셨습니다!