4강: Perceptron & Generalized Linear Model

Part 3 Generalized Linear Models

앞에서 우리는 regression과 classification문제를 다루었다. 여기서 hypothesis $h_\theta(x)$의 최적의 parameter를 찾는 방법으로 **MLE(Maximum Likelihood Estimation)**를 소개하였다. MLE의 기본 아이디어는 output y의 확률모델을 가정하고, 주어진 데이터 셋의 확률을 likelihood function으로 나타냈을 때, 이 함수의 값을 maximize하는 parameter θ를 찾는 것이다.

- linear regression에서의 확률분포 가정

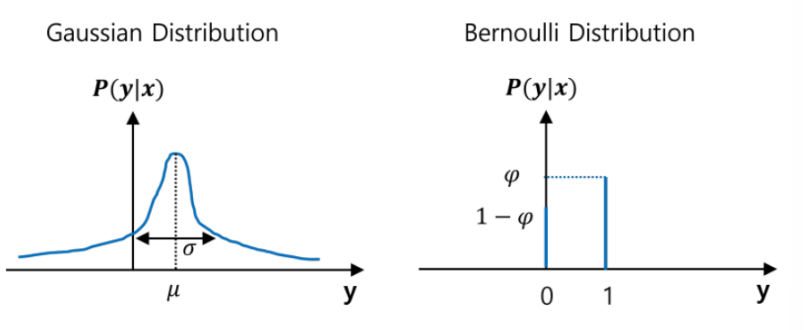
  $y|x;\theta \sim N(\mu, \sigma^2)$

  $\mu = h_\theta(x) = \theta^T x$

- logistic regression에서의 확률분포 가정

  $y|x \sim Bernoulli(\varphi)$

  $\varphi = h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{\theta^T x}}$



이러한 확률 분포를 가정함으로써 앞에서 풀었던 문제를 더 일반화 할 수 있는데, 그 시작은 위의 두가지 분포가 exponential family에 속하는 확률분포라는 것이다. 이렇게 exponential family로 확률분포를 가정한 뒤 문제를 일반화하여 나타내는 것을 GLM이라고 한다.

여기서 classification이나 regression 모두 공통적인 목적은 target variable y를 random variable x에 대한 함수 꼴로 표현하는 것이다. 이번에는 GLM(Generalized linear model)을 통해 target variable y의 성절에 따라 서로 다른 regression model을 만들어내는 과정을 살펴볼 것이다.

8. The exponential family

GLM에 대해 공부하기 전에 exponential family distribution에 대해 정리해보고자 한다.

1) Exponential family

pdf :

$$$$

- y : data
- η : natural parameter
- $T(y)$: sufficient statistic ← = y (수업에서는 t(y)가 y인 경우를 다룸,,)
- $b(y)$: base measure ← y로만 이루어진 함수
- $a(\eta)$: log partition ← η로만 이루어진 함수

각 분포들의 pmf 함수가 $T,\ b,\ a$로 나타낼 수 있으면 exponential family에 속한다는 것을 확인할 수 있다.

이제 앞서서 우리가 배운 Bernoulli와 Gaussian distribution이 exponential family에 속하는 지 확인해보자.

2) Bernoulli Distibution ($Bernoulli(\phi)$ ← ϕ : mean)

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{(1-y)}, y\ \in\ \{0,\ 1\} \\ &= exp(ylog\phi + (1-y)log(1-\phi)) \\ &= exp((log(\tfrac{\phi}{1-\phi}))y + log(1-\phi)) \end{aligned}$$

- $\eta = log(\frac{\phi}{1-\phi})$ / 흥미로운 사실은 만약 이 식을 반대로 η에 대해 나타낸다면 $\phi = \frac{1}{1+e^{-\eta}}$를 얻을 수 있다. 주목해야할 점은 ϕ가 η에 대해 시그모이드 함수라는 관계식을 갖는다는 것이다. 이것은 우연이 아닌데 이에 대해서는 GLM으로 logistic regression을 유도할 때 다시 설명할 것이다.

  ▼ 유도

$$$$

- $T(y) = y$
- $a(\eta) = -log(1-\phi) = log(1+e^{\eta})$

  ▼ 유도

  $a(\eta)$는 η에 대해 표현해야 하므로, 위에서 얻은 $\phi = \frac{1}{1+e^{-\eta}}$를 대입하면

$$\begin{aligned} a(\eta) &= -log(1-\phi) \\ &= -log(1-\tfrac{1}{1+e^{-\eta}}) \\ &= -log(\tfrac{e^{-\eta}}{1+e^{-\eta}}) \\ &= log(\tfrac{1+e^{-\eta}}{e^{-\eta}}) \\ &= log(1+e^{\eta}) \end{aligned}$$

- $b(y) = 1$

위의 과정을 통해서 Bernoulli distribution을 exponential family 형태로 나타낼 수 있다는 것을 확인할 수 있다.

3) Gaussian distribution

앞 장에서 linear regression에서 σ^2는 maximize해야 하는 term에 상수로 곱해져 있었으므로, θ와 $h_\theta(x)$의 최종 선택에 영향을 미치지 않았다는 사실을 기억하자. 따라서 아래의 유도과정을 간단히 하기 위해서 $\sigma^2 = 1$로 설정해보자.

Gaussian distribution의 pdf :

$$p(y;\mu) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(y-\mu)^2) \qquad (\sigma^2 = 1)$$

이를 exponential family의 형태 $P(y) = b(y)exp(\eta^T T(y) - a(\eta))$ 로 나타내보자.

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(y-\mu)^2) \\ &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}y^2) exp(\mu y - \frac{1}{2}\mu^2) \end{aligned}$$

- $\eta = \mu$
- $T(y) = y$
- $a(\eta) = \mu^2/2 = \eta^2/2$
- $b(y) = \frac{1}{\sqrt{2\pi}} exp(-\frac{y^2}{2})$

3) others

이 외에도 다른 많은 분포들이 exponential family에 속한다: multinomial, poisson, gamma, exponential, beta, Dirichlet 등등

exponential family에 속하는 확률 분포로 데이터의 확률 분포를 가정할 경우, GLM(Generalized Linear Model)이라는 보다 general 한 개념으로 모델 training 및 output predicting을 설명할 수 있다.

---

9. Constructing GLMs

가게의 홍보, 광고, 날씨, 요일과 같은 특정 x를 기반으로 지정된 시간 내에 가게에 도착하는 고객 수(또는 웹 사이트의 페이지 뷰 수)를 추정하는 모델을 구축하려고 하자. 우리는 이미 포아송 분포가 일반적으로 방문자 수에 대한 좋은 모델을 제공한다는 것을 알고 있다. 이때 어떻게 우리의 문제에 대한 모델을 생각해낼 수 있을까? 다행히 포아송은 exponential family distribution이므로, 일반 선형 모델(GLM)을 적용할 수 있다. 이 section에서는 이러한 문제에 대해 GLM 모델을 구성하는 방법에 대해 알아볼 것이다.

1) 개념

이전에 regression과 binary classification을 풀었던 방법을 요약해보면

1) ouput y의 확률분포를 가정한다. regression의 경우 Gaussian distribution, binary classification의 경우 Bernoulli distribution이었다.

2) Learning: 주어진 n개의 데이터 셋 $(x^{(i)}, y^{(i)})$의 확률인 likelihood functino의 값을 maximize시키는 parameter θ를 구한다.

3) Predicting:

- regression: $h_\theta(x) = \theta^T x$
- binary classification: $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

GLM의 기본개념은 output y의 확률분포로부터 시작해 모델을 학습시키고, 새로운 input에 대해 모델이 predict하는 일련의 과정을 일반화하는 것이다.

2) 가정

GLM을 도출하기 위해서는 x가 주어졌을 때 y에 대한 조건부 분포와 우리의 모델에 대해 다음의 세가지 가정을 해야한다.

1. x와 θ가 주어질 때 y는 exponential family에 속하는 확률분포를 가진다.

$y|x;\theta \sim ExponentialFamily(\eta)$

2. model의 prediction은 $T(y)$의 expected value이다. 대부분의 example에서 $T(y) = y$ 이다.

   $predict\ E[y|x]$.
   즉,
   $h_\theta = E[y|x]$

   x가 주어졌을 때 우리의 목표는 $T(y)$의 expected value를 predict하는 것이다. 대부분의 example에서 $T(y) = y$이다. 그래서 hypothesis에서 학습된 prediction $h(x)$ output이 $h(x) = E[y|x]$를 만족하기를 원한다. (이 가정은 logistic regression과 linear regression 모두에서 h_θ를 선택할 때 만족된다. 예를 들어 logistic regression에서 $h_\theta(x) = p(y = 1|x;\theta) = 0\ p(y = 0|x;\theta) + 1\ p(y = 1|x;\theta) = E[y|x;\theta]$)

3. natural parameter η와 input x는 linear한 관계를 갖는다.
   $\eta = \theta^T x \qquad \theta, x \in R^n$
   $(if\ \eta\ is\ vector-valued,\ \eta_i = \theta_i^T x)$

이제부터 logistic regression과 ordinary least squares가 GLM으로 어떻게 유도되는지 볼 것이다.

9.1 Ordinary Least Squares

y가 continuous한 값을 가질 때, y의 확률 분포를 다음과 같이 Gaussian disribution으로 가정하였다.
$y|x \sim Gaussian(\mu, \sigma^2)$

Gaussian distribution의 property에 따라 $E[y|x] = \mu$이다.
η가 아닌 μ로 parametize되어 있는 확률 분포를 exponential family form에 따라 관계식을 찾으면, $\eta = \mu$임을 밝힌적이 있다.
여기서 세번째 가정을 적용시키면, $\eta = \theta^T x$이다.

이때 두번째 가정에 의해서 $h_\theta(x)$가 ouput y의 확률분포의 expected value이므로,

$$\begin{aligned} h_\theta(x) &= E[y|x] \\ &= \mu \\ &= \eta \\ &= \theta^T x \end{aligned}$$

이다.

이것은 우리가 맨 처음 가정했던 hypothesis model과 같은 결과이다. 이것을 해석하자면, 처음에 우리는 hypothesis를 x에 대한 linear equation으로 잡고 least square error를 minimize하는 관점에서 parameter θ를 찾아 모델을 학습시켰다. 그런데, 이것은 GLM관점에서 output y가 Gaussian distribution을 갖는다는 가정으로부터 시작하여 hypothesis를 구한 것과 같은 결과를 준다는 것이다.
즉, GLM이라는 보다 일반화된 관점에서 least squre based regression문제를 설명한 결과라고 이해할 수 있다.

9.2 Logistic Regression

이제 logistic regression도 GLM의 관점에서 설명할 수 있다. binary classification에서 $y \in \{0, 1\}$이다. 이렇게 y가 binary 값으로 주어질 때 Bernoulli distribution으로 가정한다. 전 section에서 보았던 것처럼 Bernoulli distribution도 exponential family에 속한다.

$y|x;\theta \sim Bernoulli(\phi)$

$p(y|x) = \phi^y(1-\phi)^{1-y}$

이 분포는 ϕ에 대해 parameterize되어 있는데, ϕ와 η의 관계식이 $\phi = \frac{1}{1+e^{-\eta}}$ 임을 유도했었다.

이제 η가 input x에 대해 linear 하다는 관계식을 대입하면,

$\eta = \theta^T x$

$$\begin{aligned} h_\theta(x) &= E[y|x] \\ &= \phi \\ &= \frac{1}{1+e^{-\eta}} \\ &= \frac{1}{1+e^{-\theta^T x}} \end{aligned}$$

9.3 Softmax Regression(GLM)

GLM의 다른 예시에 대해 살펴보겠다. 이전까지 살펴본 binary classification에서는 ouput y는 0,1 두가지값을 가질 수 있었다. 이번에는 output y가 가질 수 있는 값을 1, 2, … k 중에서 하나의 값을 가질 수 있는 경우를 살펴볼 것이며 이는 multinomial distribution을 따른다.

1) multinomial distribution

y가 가질 수 있는 값이 k개 이므로, k개의 parameter를 사용하여 확률분포를 나타낼 수 있다.

물론 확률의 총합이 1이 되어야 하기 때문에 실제 parameter 수는 (k-1)개이고, 마지막 k번째 parameter는 나머지 parameter에 의해 결정된다. notation의 편의를 위해

$\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$로 표현할 것이다. 그러나 이것은 파라미터가 아님에 주의하자.

2) Exponential Family

이제 multinomial data를 모델링하기 위한 GLM을 유도해보자. 이를 위해서는 먼저 multinomial을 exponential family distribution형태로 표현해야 한다.

그 전에 필요한 개념과 표현법을 정리해보고자 한다.

이전의 두 예시에서는 $T(y) = y$로 주어졌는데, 여기서 $T(y)$는 상수가 아니라 (k-1)차원의 벡터로 다음과 같이 주어진다:

$$T(1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ T(2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ T(3) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ \cdots, T(k-1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \ T(k) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

이때 $T(y)_i$를 벡터 $T(y)$의 i번째 element라고 하는데, $T(y)$와 y의 관계를 다음과 같은 notation으로 나타낼 수 있다.

더 나아가 $E[T(y)_i] = p(y=i) = \phi_i$이다.

이제 드디어 multinomial가 exponential family임을 보일 준비가 되었다.

$$
\begin{aligned}
p(y;\phi) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1\{y=k\}} \\
&= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}} \\
&= \phi_1^{(T(y))_1}\phi_2^{(T(y))_2}\cdots\phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i} \\
&= \exp((T(y))_1\log(\phi_1) + (T(y))_2\log(\phi_2) + \\
&\qquad \cdots + \left(1-\textstyle\sum_{i=1}^{k-1}(T(y))_i\right)\log(\phi_k)) \\
&= \exp((T(y))_1\log(\phi_1/\phi_k) + (T(y))_2\log(\phi_2/\phi_k) + \\
&\qquad \cdots + (T(y))_{k-1}\log(\phi_{k-1}/\phi_k) + \log(\phi_k)) \\
&= b(y)\exp(\eta^T T(y) - a(\eta))
\end{aligned}
$$

where

$$
\begin{aligned}
\eta &= \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix}, \\
a(\eta) &= -\log(\phi_k) \\
b(y) &= 1.
\end{aligned}
$$

이를 통해서 우리는 mutinomial이 exponential family distribution으로 존재함을 알게 되었다…!

3) *GLM(Generalized Linear Model) & Hypothesis*

여기서 link function은 $\eta_i = log\frac{\phi_i}{\phi_k}$

또한 $\eta_k = log\frac{\phi_i}{\phi_k}$

link function을 변형하고 response function을 유도해보자

$$
\begin{aligned}
e^{\eta_i} &= \frac{\phi_i}{\phi_k} \\
\phi_k e^{\eta_i} &= \phi_i \qquad (7) \\
\phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k}\phi_i = 1
\end{aligned}
$$

이는 $\phi_k = \frac{1}{\sum_{i=1}^{k} e^{\eta_i}}$을 암시한다. 그리고 equation(7)에 대입하면 다음과 같이 **response function**을 얻을 수 있다.

$$
\phi_k = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}
$$

η와 ϕ를 연관짓는 이 함수는 **softmax function**이라고 부른다.

우리의 모델을 완성시키기 위해 η는 x와 linear 한 관계라는 세번째 가정을 사용해보자. 따라서 $\eta_i = \theta_i^T x (for\ i = 1, \ldots, k-1),\ where \theta_1, \ldots, \theta_{k-1} \in R^{d+1}$ 은 우리 모델의 파라미터이다.

notation의 편의성을 위해, $\theta_k = 0$으로 정의한다. 그래서 $\eta_k = \theta_k^T x = 0$이다.

그러므로 우리 모델의 확률 분포는 다음과 같다.

$$\begin{aligned} p(y=i|x;\theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \end{aligned} \tag{8}$$

$y \in \{1, \ldots, k\}$의 분류 문제에 적용되는 이 모델을 **softmax regression**이라고 부른다. 이것은 logistic regression의 일반화된 모델이다.

그래서 이것의 hypothesis는 다음과 같다.

$$\begin{aligned} h_\theta(x) &= \mathrm{E}[T(y)|x;\theta] \\ &= \mathrm{E}\left[\left. \begin{array}{c} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{array} \right| x;\theta \right] \\ &= \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\exp(\theta_1^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \\ \frac{\exp(\theta_2^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \\ \vdots \\ \frac{\exp(\theta_{k-1}^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \end{bmatrix}. \end{aligned}$$

마지막으로 parameter fitting에 대해 알아보자. 우리가 n개의 training data가 있고 모델의 parameter θ에 대해 학습하고 싶다면 다음과 같이 loglikehood function을 쓸 수 있다.

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \log \prod_{l=1}^{k} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)}=l\}} \end{aligned}$$

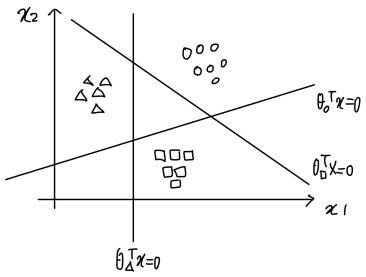
gradient ascent나 newton's method와 같은 방법을 이용해서 $l(\theta)$를 최대화하는 θ를 구하면 된다.

9.3 softmax Regression (non-GLM)

강의노트에는 GLM 접근방식으로 softmax regression을 유도했는데 강의에서는 다른 접근 방식으로 유도하여 이를 추가적으로 설명하고자 한다.

이는 cross entropy를 최소화하는 방식으로 파라미터를 구하는 방법이다.

다음과 같은 k개의 그룹이 있는 muticlassification 문제가 있다고 하자.



k : # of classes

$x^{(i)} \in R^n$

label $y = \{\{0,\ 1\}^k\}$ $\quad e.g.\ [0,\ 0,\ 1,\ 0]$

이렇게 class를 one hot vector로 표현할 수도 있다.

softmax regression

each class has its own set of parameters

$\theta_{class} \in R^n$