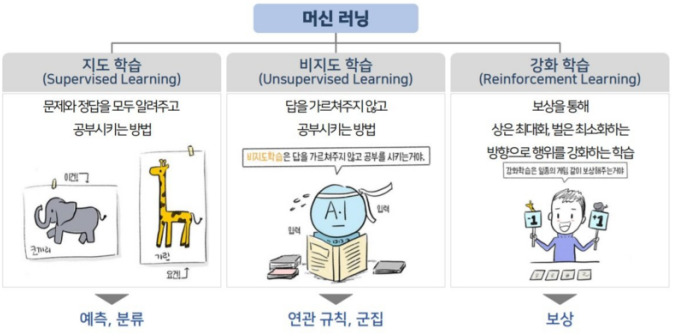
CH1: Motivations and Basics

1.1 Motivation

머신 러닝 (Machine Learning)은 학습하는 방법, 즉 데이터를 최대한 활용해 특정 작업의 성능을 향상시키는 방법을 이해하고 구현하는 탐구 분야이다. 머신 러닝은 통계, 데이터 베이스, 데이터마이닝, 산업 분야등 다양한 분야에 걸쳐있으며, 데이터를 얻을 수 있다면 모든 분야에 적용이 가능하다.

머신러닝의 종류는 다음과 같다.



지도학습(Supervised Learning)이란, 데이터에 대한 학습데이터가 주어진 상태에서 컴퓨터를 학습시키는 방법이다. 쉽게 말하면, Y값이 주어진 경우라고 생각하면 된다.

비지도학습(Unsupervised Learning)이란, 데이터에 대한 학습데이터가 없는 상태에서 오직 입력데이터만 이용해서 컴퓨터를 학습시키는 방법이다. 이에 대한 예시로는 clustering, LDA 등이 있다.

1.2 Warming Up With A Short Episode

이 부분을 시작하기 앞서 예시를 하나 들어보자. 압정을 던저 머리(H)와 꼬리(T)를 베팅하는 도박을 하고 있다. 이때, 핀이 위로 올라가면, 판돈 2배를 준다. 압정은, 모양이 균일하지 않기 때문에 머리와 꼬리가 나올 확률이 50:50으로 동일하다고 할 수 없다.

어떤 부자가 압정을 던져서 나온 결과를 분석해보라고 제안을 했다고 해보자. 부자가 “어떻게 하면 내가 돈을 딸 수 있을까?라고 물어보았다. 나는 우선 ‘몇 번 던져보세요’라고 대답했다. 부자는 5번정도 압정을 던졌는데, 머리 2번, 꼬리 3번이 나왔다. 다음과 같은 결과를 토대로 꼬리가 더 많이 나왔으니 꼬리에 배팅하고 말할 수 있을까?

위의 물음에 대한 답을 하려면 확률에 대한 지식이 필요하다. 우선 위의 실험을 이항분포 (Binomial Distribution)이라는 확률분포를 통해 계산해보자.

1.2.1 Binomial Distribution

이항분포 (Binomial Distribution)란, 연속된 n번 독립적 시행에서 각 시행의 확률이 θ를 가지는 경우이다. 이는 베르누이 시행을 n번 반복한 것과 동일하다. 이항분포의 확률질량함수는 다음과 같다.

$$f(k;n;p) = P(K=k) = \binom{n}{K} p^k (1-p)^{n-k}\ ,\ k = 0, 1, ..., n$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

이항 분포에서 확인하는 일은 k번의 성공과 n-k번의 실패를 했을 확률이다. 이때, 이항분포는 k값이 0부터 n까지 모든 경우의 수에 대한 확률을 조사해 각각의 경우에 대한 확률을 미리 계산해둔 것이라고 할 수 있다. 이때, 각 회차는 i.i.d (independetly identically distributed), 즉, 서로 독립적이고 동일하게 분포되어 있다고 가정한다.

다시 예시로 돌아가보자면, 압정을 던지는 각각의 시행(Event)는 서로 독립이다. 그러면 이제 압정을 던지는 event를 확률로 표현해보자.

- $P(H) = \theta$: 압정을 던졌을 때 Head가 나올 확률
- $P(T) = 1 - \theta$: 압정을 던졌을 때 Tail이 나올 확률

앞의 예시를 확률로 표현하면 다음과 같이 표현할 수 있다.

$$P(HHTHT) = \theta\theta(1-\theta)\theta(1-\theta) = \theta^3(1-\theta)^2$$

1.2.2 Maximum Likelihood Estimation (MLE)

만약 압정을 던진 결과가 다음과 같은 이항분포를 따른다고 가정하자, $P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$. $P(D|\theta)$는 θ라는 확률정보가 주어졌을 때 D라고하는 데이터가 관측될 확률을 뜻한다. 이때, θ를 최대화 할 수 있는 방법은 무엇이 있을까?

Maximum Likelihood Estimation (MLE)는 모수적인 데이터 밀도 추정 방법으로, 파라미터 $\theta = (\theta_1, ..., \theta_m)$로 구성된 어떤 확률밀도함수 $P(x|\theta)$에서 관측된 표본 데이터 집합을 $x = (x_1, x_2, ..., x_n)$이라고 할 때, 이 표본들에서 파라미터 θ를 추정하는 방법이다. 즉, 어떤 확률변수에서 표집한 값들을 토대로 그 확률변수의 모수를 구하는 방법이며, 어떤 모수가 주어졌을 때, 원하는 값들이 나올 가능도(likelihood)를 최대로 만드는 모수를 선택하는 방법이다. 결국 MLE는 Likelihood 함수의 최대값을 찾는 방법이다. 이는 점추정 방식에 속하며 가능도는 다음과 같다.

$$L(\theta) = f_\theta(x_1, x_2, ..., x_n)$$

여기서, 가능도를 최대로 만드는 θ는 다음과 같다.

$$\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta P(D|\theta)$$

이를 우리의 상황에 적용하면, $\hat{\theta} = \arg\max_\theta P(D|\theta) = \arg\max_\theta \theta^{a_H}(1-\theta)^{a_T}$ 으로 쓸수 있다.

log 함수는 단조증가 함수이기 때문에 likelihood function의 최대값을 찾으나, log-likelihood function의 최대값을 찾으나, 두 경우 모두 최대값을 갖게 해주는 정의역의 함수 입력값은 동일하다. 따라서, 보통은 계산의 편의를 위해 log-likelihood의 최대값을 찾는다. 이에 대한 수식은 다음과 같이 나타낼 수 있다.

$$\hat{\theta} = \arg\max_\theta \ln P(D|\theta) = \arg\max_\theta \ln[\theta^{a_H}(1-\theta)^{a_T}] = \arg\max_\theta [a_H \ln\theta + a_T \ln(1-\theta)]$$

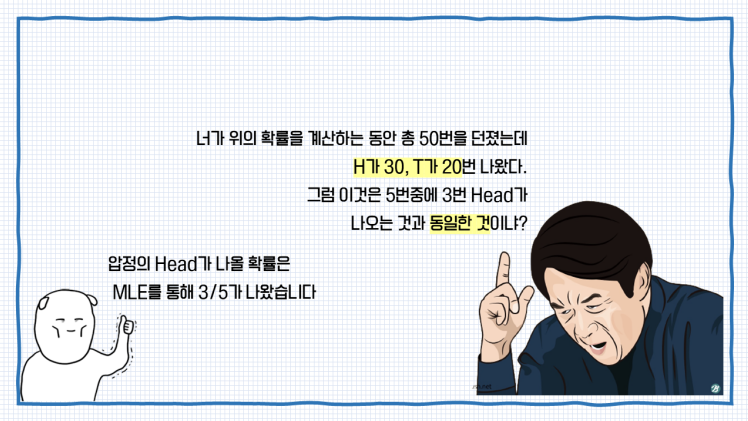
어떤 함수의 최대값을 찾는 방법 중, 가장 보편적인 방법은 미분계수를 이용하는 것이다. 즉, 찾고자 하는 파라미터 θ에 대해 편미분하고, 그 값이 0이 되도록 하는 θ를 찾는 과정을 통해 likelihood 함수를 최대화 할 수 있다.

$$\begin{aligned} \frac{d}{d\theta}(a_H \ln\theta + a_T \ln(1-\theta)) &= 0 \\ \frac{a_H}{\theta} - \frac{a_T}{1-\theta} &= 0 \\ \theta &= \frac{a_H}{a_T + a_H} \end{aligned}$$

따라서,θ는 다음과 같은 상황에서 최대화 되며, MLE는 $\hat{\theta} = \frac{a_H}{a_T + a_H}$이다.

1.2.3 Simple Error Bound

이제 '압정이 Head가 나올 확률은 MLE를 통해 3/5가 나왔다'라고 부자에게 말할 수 있다. 부자는 '그래 좋다. 너가 위의 확률을 계산하는 동안 총 50번을 던졌는데 H가 30, T가 20번 나왔다. 그럼 이것은 5번중에 3번 Head가 나오는 것과 동일한 것이냐?'라고 물었다. 과연 두 가지 결과는 같다고 말할 수 있을까?



질문에 대한 답은 '아니다'이다. 일단, 우리가 지금까지 알아본 $\hat{\theta}$은 그저 추정값일 뿐, 실제 확률인 θ^*가 아니다. 추정값은 언제나 실제 값과 오차가 있기 마련이다. 즉, $\hat{\theta} \neq \theta^*$이다. 오차는 둘 사이의 차이값으므로 절댓값을 이용하여 $|\hat{\theta} - \theta^*|$로 나타낼 수 있다. 오차범위를 나타내는 방법은 많은데, Hoeffding's inequailty에 따르면, 다음과 같이 나타낼 수 있다.

$$P(|\hat{\theta} - \theta^*|) \leq 2e^{-2N\epsilon^2} \;\; where \; N = a_H + a_T$$

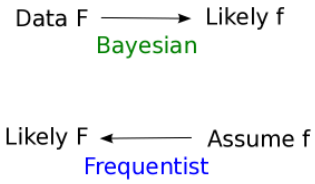
위의 식을 보면, 오차가 임의의 작은 값ϵ보다 커질 확률은 $2e^{-2N\epsilon^2}$으로 나타난다. 즉, ϵ이 동일한 조건에서 실행횟수 N이 증가할수록 오차의 범위가 줄어들게 된다는 의미이다. 이러한 학습 방식은 PAC 학습 (Probably Approximate Correct learning, PAC learning)이라고 한다. PAC learning의 목적은 높은 확률로 낮은 오차 범위를 갖도록 하는 것으로, 이를 달성하기 위해서는 큰 데이터셋을 가지고 있어야 한다.

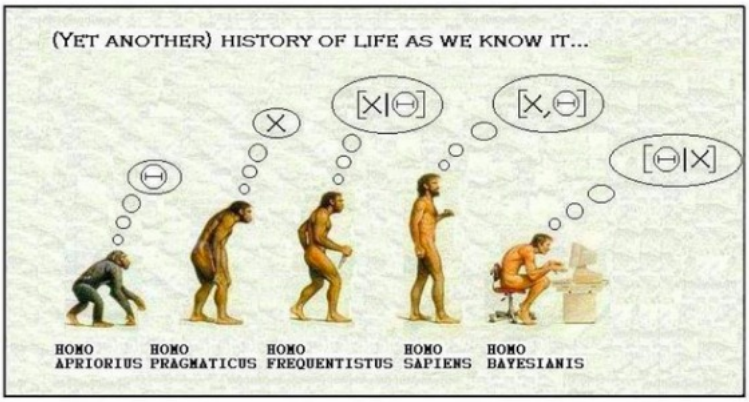
~~알아서 읽어보기 ^^~~

hoeffding inequality.pdf

1.2.4 Maximum a Posteriori Estimation (MAP)

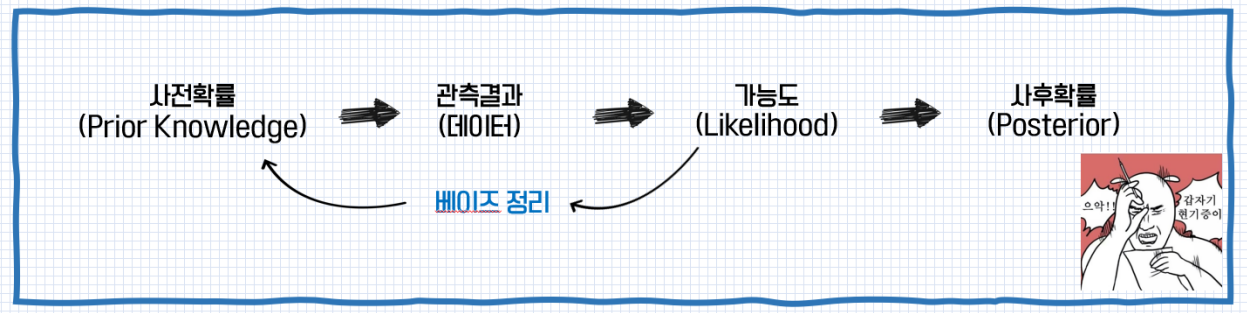
모든 것이 잘 해결되어가는 이때, 갑자기 베이즈가 (수학자겸 신학자임, 처음으로 귀납적 확률을 사용한 사람임) 부자에게 '압정을 던지면 정말 60%의 확률로 head가 나올거 같냐..? Head/Tail이 나올확률이 50:50이라고 생각안하세요…?'라고 물었다. 부자는 '나도 처음에는 50:50의 확률로 나올것 같다고 생각했는데, 막상 던져보니까 60%가 나오더라.'라고 대답했다. 이에 베이즈는 '압정의 Head와 Tail이 50:50의 확률로 나온다는 사전정보를 parameter 추정되정에 넣을 수 있습니다! 사전정보를 내포할 수 있는 $P(\theta)$에 대해 한 번 알아보죠!'





통계학은 총 2가지 관점으로 나뉘는데, 하나는 확률을 사건의 빈도로 보는 것인 frequentist와 확률을 사건 발생에 대한 믿음 또는 척도로 바라보는 관점인 Baysian이 있다. 이 두 학파들은 '확률을 해석하는 관점의 차이'라고 설명할 수 있다.

베이지안 방법은 수학적 배경이 까다롭고, 계산량이 많기 때문에 구현의 어려움이 있어, 예전에는 통계학자들로부터 환영받지 못했다. 하지만 컴퓨터의 연산 능력확장과 다양한 알고리즘 개발로 인해 베이지안 방법도 통계/머신러닝에서 많이 사용되고 있다. 베이지안의 확률추론 방법은 다음과 같다.



여기서 베이즈 정리란 다음과 같다.

$$Posterior = P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{Likelihood \times Prior\ Knowledge}{Normalizing\ Constant}$$

- $P(D)$: 데이터가 관측될 확률로 이미 주어진 사실이기 때문에 우리가 어떻게 할 수 없다. 그래서 Noramlizing Constant로 생각하며, 계산시에는 proportion out 시킨다.
- $P(\theta)$: 사전정보로 latent한 drive force이기 떄문에 중요하다
- $P(D|\theta)$: θ가 주어졌을때 데이터를 관측할 likelihood
- $P(\theta|D)$: 사후확률로, 데이터가 주어졌을 때 θ의 확률을 의미한다

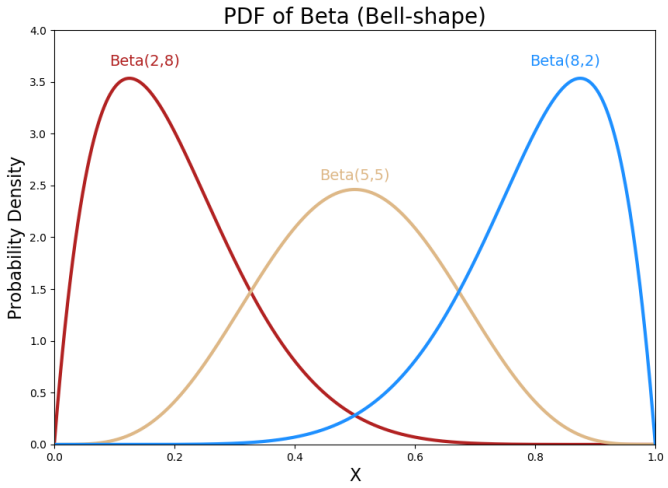
실제 데이터를 처리하는데 있어서 베이즈 규칙을 적용하기 위해서는 수학적인 테크닉이 필요하다. 특히, 사전확률의 분포와 사후분포와 사후확률의 분포와 서로 밀접한 관계가 있다면, 반복적으로 베이즈 규칙을 적용하는데 용이할 것이다. 즉, 추가적으로 데이터를 넣으면서, 파라미터 θ를 지속적으로 업데이트 할 수 있다. 사전확률에 가능도를 곱하여도, 사전확률의 분포와 사후확률의 분포가 같은 형태가 될 때, 이 사전분포 $P(\theta)$를 가능도 $P(Y|\theta)$에 대한 켤레 사전분포 (conjugate prior)라고 한다. 켤레 사전분포는 특정 가능도 함수에만 적용되며, 사후 분포를 간단하게 수학적으로 표현할 수 있다는 장점이 있다.

우리의 예시에 사용된 분포는 이항분포였는데, 베르누이(이항분포) 가능도 함수에 대한 켤레 사전분포는 베타 분포이다. 베타분포의 밀도 함수는 다음과 같으며, 이때 α, β는 N번 시행했을 때 앞면이 나올 사건 α, 뒷면이 나올 사건 β를 뜻한다.

$$f(\theta;\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \quad , \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \rightarrow Beta(\alpha,\beta)$$

베타분포의 특징은 다음과 같다.

- α값이 커질수록 θ값이 오른쪽으로 기울며 θ값은 점점 커진다. 반대로 β값이 커지면 왼쪽으로 치우친다. α, β가 함께 커질수록 베타분포의 모양이 좁아진다.
- $\alpha = \beta = 1$인 경우, θ가 속하는 [0,1]사이에서 $P(\theta)$의 값이 uniform distribution의 형태를 띈다.



이제 사후 분포를 구해보자!

$$P(\theta|D) \propto P(D|\theta)P(\theta) \propto \theta^{x}(1-\theta)^{N-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{x+\alpha-1}(1-\theta)^{N-x+\beta-1} \sim Beta(x+\alpha, N-x+\beta)$$

이 수식을 우리의 예시에 적용해보면 $P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1} \sim Beta(a_H+\alpha, a_T+\beta)$이 된다.

앞서 MLE는$P(D|\theta)$를 이용해 $\hat{\theta}$를 구할 수 있었는데, MAP는 $P(\theta|D)$를 통해 $\hat{\theta}$를 구하게 된다.

MAP(Maximum a Posterior/Posteriori)는 주어진 관측결과와 사전지식을 결합해서 최적의 모수를 찾아내는 방법이다. MAP는 MLE와 마찬가지로 점 추정을 사용할수 있지만, MLE는 어떤 사건이 일어날 확률을 가장 높이는 모수를 찾는 것에 비해, MAP는 모수의 사전 확률과 결합된 확률을 고려한다는 점이 다르다. MLE와 동일하게 극점을 이용한 최적화를 하면 다음과 같다.

$$\hat{\theta} = \arg\max_{\theta} P(\theta|D)$$
$$P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{\alpha_T+\beta-1}$$
$$\hat{\theta} = \frac{a_H+\alpha-1}{\alpha_H+\alpha+\alpha_T+\beta-2}$$

위의 식을 이용하면, 사전 정보를 넣어 $\hat{\theta}$를 조절할 수 있다.

최대 우도 추정과 최대 사후 확률 추정을 통해 추정한 특정 사건의 확률 θ는 비슷한 모양을 띠고 있지만 완전히 같지는 않다. 베타 분포 내의 상수가 $\alpha = \beta = 1$일 경우에만 서로 같아진다. 하지만, 시행 횟수 N이 커지면, 즉, 데이터셋의 크기가 커지면, 상수인 α, β에 비해 α_H, α_T가 커지게 된다. 그렇기 때문에, N이 커질수록 최대 우도 추정으로 구한 θ와 최대 사후 확률 추정으로 구한 θ가 같아진다. 반대의 경우에서는 사전정보가 중요한 역할을 한다.

1.3 Basics

1.3.1 정규분포

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{x-\mu)^2}{2\sigma^2}} \rightarrow N(\mu,\sigma^2)$$

평균: μ / 분산: σ^2

1.3.2 베타분포

$$f(\theta;\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \quad , \quad B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \rightarrow Beta(\alpha,\beta)$$

평균: $\frac{\alpha}{\alpha+\beta}$ / 분산: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

1.3.3 다항분포 (Multinomial Distribution)

$$f(x_1,...,x_k;n,p_1,...,p_k)=\frac{n!}{x_1!...x_k!}p_1^{x_1}...p_k^{x_k} \quad \rightarrow Mult(p)$$

평균: np_i / 분산: $np_i(1-p_i)$