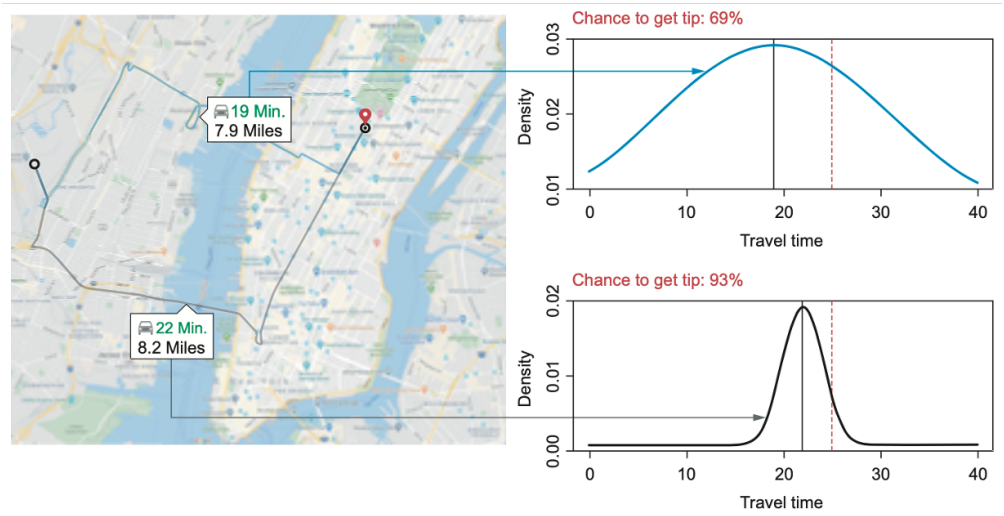
ESC 2024 Winter Session 1st Week

Curve Fitting, Decision Theory, & Information Theory with Probability

전인태

1. Probability on Machine Learning

Motivation : 우리는 택시기사이며, 급한 약속이 있는 승객이 목적지까지 25분 내에 도착한다면 돈을 두배로 주겠다는 제안을 했다고 하자. 네비게이션에 목적지를 입력하였더니 두 가지 경로를 알려주었다. 그러나 실제로는 19분이 걸리는 경로가 자주 막히는 길일 수도 있으며, 평균적으로 조금 더 오래 걸리더라도 변동성이 작은 22분 아래 경로를 참고하는 것이 더 나은 선택일 것이다. 즉, 확률을 고려한다면 더 효과적으로 문제를 해결할 수 있다.



이때 확률을 도입하기 위한 방법으로 통계적 방법론을 적용할 수 있으며, Frequentist viewpoint와 Bayesian viewpoint를 모두 사용할 수 있다. 이 중 Bayesian viewpoint는 모델로 하여금 uncertainty를 고려할 수 있기 때문에 특정 상황에서 유용하게 사용될 수 있다.

2. Probabilistic & Non-probabilistic Curve Fitting

$(\mathbf{x}, \mathbf{t})$를 모델을 학습시키기 위한 training dataset이라 하자. 각각 input data와, 이에 대응되는 target data를 의미한다. M차 polynomial regression model

$$y(x, \mathbf{W}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

을 데이터를 표현하기 위한 모델로 잡았을 때, 실제 target과의 오차를 수치화하기 위하여 loss function을 정의할 수 있다. 이때 loss function을 최소화하는 $\mathbf{w}$를 찾아 최적화된 model을 만들 수 있으며, loss

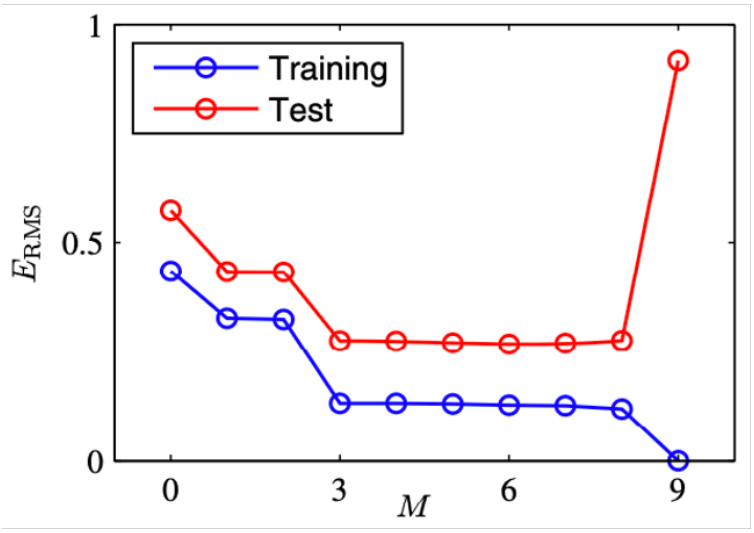
function으로서 가장 많이 쓰이는 것은 에러의 제곱합(SSE)이다 :

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2.$$

M을 결정하기 위하여 RMS (root-mean-square) error를 도입하자 :

$$E_{RMS} = \sqrt{E(\mathbf{W}^*)/N}.$$

N으로 나누는 것은 서로 다른 데이터 개수를 가진 모델들을 비교할 수 있게 해주며, $\sqrt{\cdot}$은 error가 t와 동일한 스케일에서 측정될 수 있게 해준다. ($\mathbf{w}^*$은 $\text{argmin}_{\mathbf{w}}\, E(\mathbf{w})$을 의미한다.) 이때 특정한 M에서 RMS값이 폭증할 때가 있는데, 이는 곧 모델이 데이터에 과적합되었음을 나타낸다.



과적합을 방지하기 위해서는 training dataset의 데이터 개수를 늘리거나, regularization을 이용할 수 있다. Regularization은 penalty term을 추가함으로써 계수가 너무 커지는 것을 방지해주며, penalty term의 norm으로 어떤 것을 사용하느냐에 따라 **Ridge**(L2 norm), **Lasso**(L1 norm), **Elastic Net**(L1+L2)으로 구분된다.

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

위와 같은 ML적 관점이 아닌, 전통적인 통계학 방식으로 접근해보자. $\mathbf{W}_{ML}$를 찾기 위하여 MLE를 이용할 수 있다. 이러한 방법의 장점은 $p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = N(t|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1}$이라는 predictive distribution을 찾음으로써, 단순한 point prediction이 아닌, 예측되는 target variable의 분포까지 파악할 수 있다는 것이다.
마찬가지로 또 다른 통계적 방법인 Bayesian 관점을 이용하여 MAP를 찾는다고 하더라도, 위와 동일하게 predictive distribution을 찾을 수 있다. 여기서 주목해야 할 것은, $p(t|x, \mathbf{x}, \mathbf{t}) \sim N(t|m(x), s^2(x))$라는 분포의 평균과 분산이 test point인 x에 의존한다는 것이다. 즉, Bayesian은 test point의 uncertainty를 반영하고 있음을 보여준다.

3. Decision Theory

각 input space를 **decision region**이라 불리는 R_k로 나누고, 각 class 당 하나의 region을 할당하자. 만약 어떤 데이터 $x \in R_1$에 대하여, $p(C_2|x) >> p(C_1|x)$라면 x는 R_1에 속함에도 C_1이 아닌 C_2에 속하게

될 것이다. 이때 misclassification이 발생하게 되며, 이러한 일이 발생하지 않도록 p(misclassification)이 최소화되게끔 decision region을 정하는, 즉, decision boundary를 정하는 것이 decision theory의 목표이다.
또, 모델의 판단 오류가 큰 위험을 불러일으킬 수 있는 경우, 오류 확률이 클 때 판단을 유보하도록 할 수 있는데, 이를 **reject option**이라 하며, 이는 reject 여부를 결정하는 임계값 θ를 설정하여 구현할 수 있다. 만약 $\theta = 1$이라면 어떤 k에 대해서든 $\max_k\{p(C_k|x)\} \leq \theta = 1$이 되므로, 모든 판단이 이루어지지 않을 것이며, $\sum k = K$일 때, $\theta < 1/K$라면 어떤 데이터도 rejected 되지 않을 것이다.

4. Information Theory

우리가 어떤 정보를 전송한다 하자. 정보의 양은 발생할 것으로 예상되는 사건일수록 작으며, 예상되지 않을수록 커진다. 어떤 discrete random variable $x \sim p(x)$에 대하여, x의 정보량 $h(x)$는

$$h(x) = -\log_2 p(x)$$

로 정의되며, 단위는 **bits**이다. 이때 정보의 평균량 $E[h(x)]$는 $H[x]$라 표기하며,

$$H[x] = -\sum_x p(x)\log_2 p(x)$$

로 정의되고, 확률변수 x의 **entropy**라 불린다. 또한 $\lim_{p\to 0} p\ln p = 0$이므로 $x = 0$일 때의 $p(x)\ln p(x) = 0$으로 한다. $H[x]$는 $p(x_i) = p(x_j)\ \forall i \neq j$일 때 최대가 되며, $p(x_k) = 1;$ o.w. 0일 때 최소가 된다.
만약 x가 continuous random variable이라면,

$$H[x] = -\int p(x)\ln p(x)dx$$

으로 정의되며 differential entropy라 부른다. 하지만 연속적인 데이터를 이산적으로 전송되는 디지털 체계에서 완벽하게 전송하는 것은 사실상 불가능하다. 따라서 연속적인 데이터를 Δ 만큼의 정보로 쪼개어 이산적으로 변형하고,

$$H_\Delta = -\sum_i p(x_i)\Delta\ln\left(p(x_i)\Delta\right) = -\sum_i p(x_i)\Delta\ln p(x_i) - \ln\Delta$$

와 같이 전송하는 것이 일반적이다. 여기에서 $\ln\Delta$는 연속적인 데이터를 이산적으로 변형하여 전송할 때의 오류로, $\Delta \to 0$에 따라 $\ln\Delta$가 diverge 하는 것에서 연속적인 데이터를 이산적으로 전송되는 디지털 체계에서 완벽하게 전송하는 것은 사실상 불가능하다는 것을 수식적으로 확인할 수 있다. Differential entropy는 $p(x)\ N(\mu, \sigma^2)$일 때 최대가 된다.
우리가 알지 못하는 확률분포 $p(\mathbf{x})$를 전송하기 위하여, 근사 확률분포인 $q(\mathbf{x})$를 찾고, 이를 전송할 수 있다. 이때 p와 q 간의 오차에 의하여, 추가적인 정보량이 필요한데, 이때 필요한 평균 추가 정보량을

$$\begin{aligned} KL(p||q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})d\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}d\mathbf{x} \end{aligned}$$

로 정의하며, **relative entropy**, **Kullback-Leibler divergence**, 혹은 **KL divergence**라 부른다. 이때, $KL(p||q) \neq KL(q||p)$임에 유의하자.