6. Probability and Distributions

Probability: Fraction of items an event occurs, or as a degree of belief about an event

To measure the chance of sth occurring in an experiment

Quantifying uncertainty requires the idea of a **random variable**

Probability distribution: A function that measures the probability that a particular outcome will occur

6.1 Construction of a Probability Space

Theory of probability: A math. Structure to describe random outcomes of experiments

Goal is to perform automated reasoning

Baysian interpretation usese probability to specify the degree of uncertainty that the user has about an event.

Subjective probability or degree of belief.

6.1.2 Probability and Random Variables

Sample space(표본 공간): 시행(Experiment)의 모든 가능한 결과의 집합

Event space(샘플 공간): 모든 가능한 event 들의 집합

The probability(확률)

Event(사건): Sample space 의 부분 집합

6.1.3 Statistics

Probability: Consider a model of some process(Underlying uncertainty is captured by random varaibles, and we use the rules of probability to derive what happens)

Statistics: observe that sth happened and try to figure out the underlying process that explains the observations

ML is close to statistics

Another aspect of ML systems: we are interested in **generalization error**

Interested in performance of our system on instances that we will observe in the **future**

6.2 Discrete and Continuous Probabilities

Whether target space is discrete or continuous, the natural way to refer to distributions is different

Discrete rv -> probability mass function PMF

Continuous rv -> cumulative distribution function CDF

6.2.1 Discrete Probabilities

Target space of the joint probability = Cartesian product of the target spaces of each of the rv

*데카르트의 곱: A * B = {(x, y) | x∈A and y∈B}

Joint probability: Probability of the intersection of both events

In ML, use discrete probability distributions to model **categorical variable**

Ex. Variables that take **a finite set of unordered values**

Can be **categorical features** such as the degree taken at university when used for predicting the salary of a person

Categorical labels such as letters of the alphabets

6.2.2 Continuous Probabilities

Real-valued rv

Probability spaces with finite states X <= 1. Repeat sth infinitely often (When discuss generalization errors)

*Generalization error(일반화의 오류): 알고리즘이 전례가 없는 데이터에 대한 결과값을 얼마나 정확하게 예측할 수 있는지의 정도.

2. Want to draw a point from an interval (Continuous distributions such as the Gaussian)

*Gaussian distribution(가우스 분포)=정규 분포: 연속확률분포의 하나. 수집된 자료의 분포를 근사하는 데 자주 사용되며, 이것은 중심극한정리에 의하여 독립적인 확률 변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문

-Probability density function(확률 밀도 함수, PDF) used to describe continuous probability distributions

-Cumulative Distribution Function(누적분포함수, CDF): 어떤 확률 분포에 대해서 확률 변수가 특정 값보다 작거나 같은 확률

F(x) = P(X<=x)

CDF 미분하면 PDF, PDF 적분하면 CDF

6.2.3 Contrasting Discrete and Continuous Distributions

For discrete rv, probability of each state must lie in the interval [0, 1]

For continuous rv, the normalization does not imply that the value of the density is less than or equal to 1 for all values

Uniform distribution(연속균등분포) for both discrete and continuous random variables

Type	"Point probability"	"Interval probability"
Discrete	$P(X = x)$ Probability mass function	Not applicable
Continuous	$p(x)$ Probability density function	$P(X \leqslant x)$ Cumulative distribution function

6.3 Sum Rule, Product Rule, and Bayes' Theorem

1. sum rule

2. product rule

Bayes' Theorem(베이즈 정리)

conclusions about $\boldsymbol{x}$ given the observed values of $\boldsymbol{y}$. *Bayes' theorem* (also *Bayes' rule* or *Bayes' law*)

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}} \tag{6.23}$$

is a direct consequence of the product rule in (6.22) since

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y}) \tag{6.24}$$

and

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \tag{6.25}$$

so that

$$p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \iff p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x})}{p(\boldsymbol{y})}. \tag{6.26}$$

6.4 Summary Statistics and Independence

Means

Covariance: Expected product of their deviations from their respective means

두 개 또는 그 이상의 랜덤 변수에 대한 의존성. 두 도메인의 객체가 서로 어떤 영향을 끼치존가를 측정하는 지표.

Cov[x, y] = E[xy] - E[x]E[y]

Covariance matrix: tell about the spread of the data

Expected value of a function

기댓값. 어떤 확률을 가진 사건을 무한히 반복했을 경우 얻을 수 있는 값의 평균으로서 기대할 수 있는 값

Correlation(상관분석): A 변수가 증가함에 따라 B 변수도 증가/감소하는지 분석. 두 변수 사이의 선형적인 관계 정도를 나타내기 위해 상관계수(correlation coefficient)를 사용함.

6.4.2 Empirical Means and Covariances

Empirical statistics

1. Finite dataset

2. Pbserve the data

Empirical mean or sample mean

6.4.3 Three Expressions for the Variance

Standard definition of variance: Expectation of the squared deviation of a rv X from its expected value

The mean of the square minus of the square of the mean

A sum of pairwise differences between all pairs of observations

6.4.4 Sums and Transformations of Random Variables

Consider two random variables X, Y with states $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$. Then:

$$\mathbb{E}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{y}] \tag{6.46}$$

$$\mathbb{E}[\boldsymbol{x} - \boldsymbol{y}] = \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{y}] \tag{6.47}$$

$$\mathbb{V}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}] \tag{6.48}$$

$$\mathbb{V}[\boldsymbol{x} - \boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]\,. \tag{6.49}$$

Intuitively, two random variables X and Y are independent if the value of $\boldsymbol{y}$ (once known) does not add any additional information about $\boldsymbol{x}$ (and vice versa). If X, Y are (statistically) independent, then

- $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
- $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
- $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$
- $\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$

The last point may not hold in converse, i.e., two random variables can have covariance zero but are not statistically independent. To understand why, recall that covariance measures only linear dependence. Therefore, random variables that are nonlinearly dependent could have covariance zero.

6.4.5 Statistical Independence

Two rv X, Y are statistically independent if and only if

P(x, y) = p(x)p(y)

Definition 6.10 (Independence). Two random variables X, Y are *statistically independent* if and only if

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})\,. \tag{6.53}$$

Intuitively, two random variables X and Y are independent if the value of $\boldsymbol{y}$ (once known) does not add any additional information about $\boldsymbol{x}$ (and vice versa). If X, Y are (statistically) independent, then

- $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
- $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
- $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$
- $\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$

Conditional independence