NAÏVE BAYES CLASSIFIER

Week07

classification

Naïve Bayes Classifier

Review: Types of Classifiers

- A classifier is a function that assigns to a sample, $\mathbf{x}$ a class label $\hat{y}$

$$\hat{y} = f(\mathbf{x})$$

- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \boldsymbol{X}$, they assign probabilities to all $y \in Y$
  - Hard classification

$$\hat{y} = \arg\max_{y} \Pr(Y = y|\,\mathbf{x})$$

- Logistic regression

$$f(x) = P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p}}$$

Any other ways to model $P(Y|X)$?

Naïve Bayes Classifier

- Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumption between features
  - Bayes' theorem

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

    - A and B are events
    - $P(A)$ and $P(B)$ are the probabilities of A and B without regard to each other
    - $P(A|B)$, a conditional probability, is the probability of A given that B is true
    - $P(B|A)$, is the probability of B given that A is true

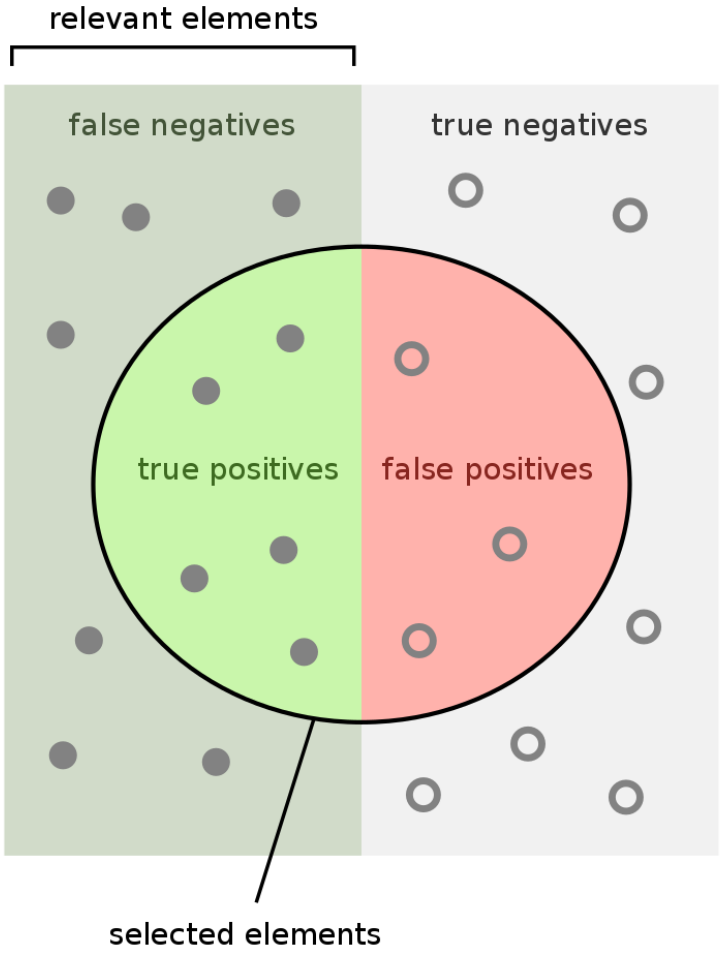
Example of Bayes' Theorem

- Suppose a drug test is 99% sensitive and 99% specific
  - 99% sensitive=99% true positive over real positive
  - 99% specific=99% true negative over real negative

Decision \ Real	Positive	Negative
Positive	True positive	False positive (Type I error)
Negative	False negative (Type II error)	True negative

- Suppose that 0.5% of people are users of the drug

※ Sensitivity and Specificity



How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

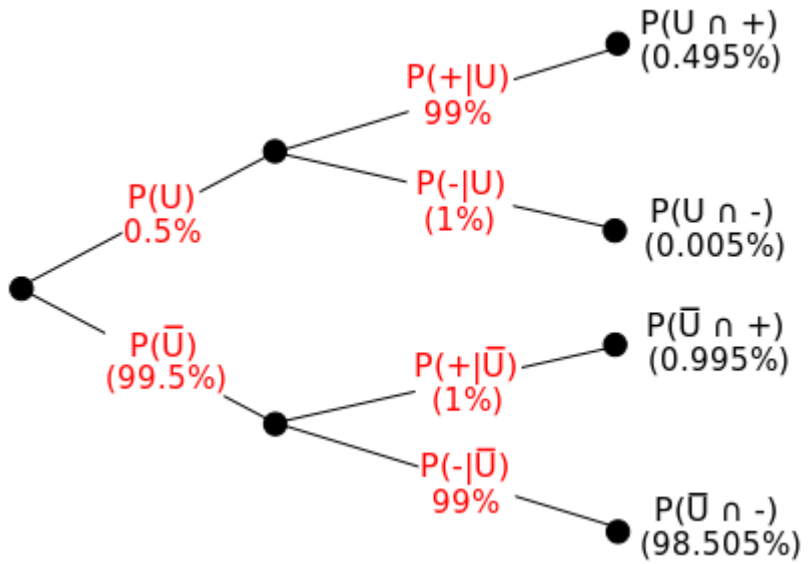
How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

〈https://en.wikipedia.org/wiki/Sensitivity_and_specificity〉

Example of Bayes' Theorem

- If a randomly selected individual tests positive, what is the probability he or she is a user of drug?
  - This problem is to calculate $P(U|+)$
    - $+$ means positive drug test
    - U represents user, $\bar{U}$ represents non-user



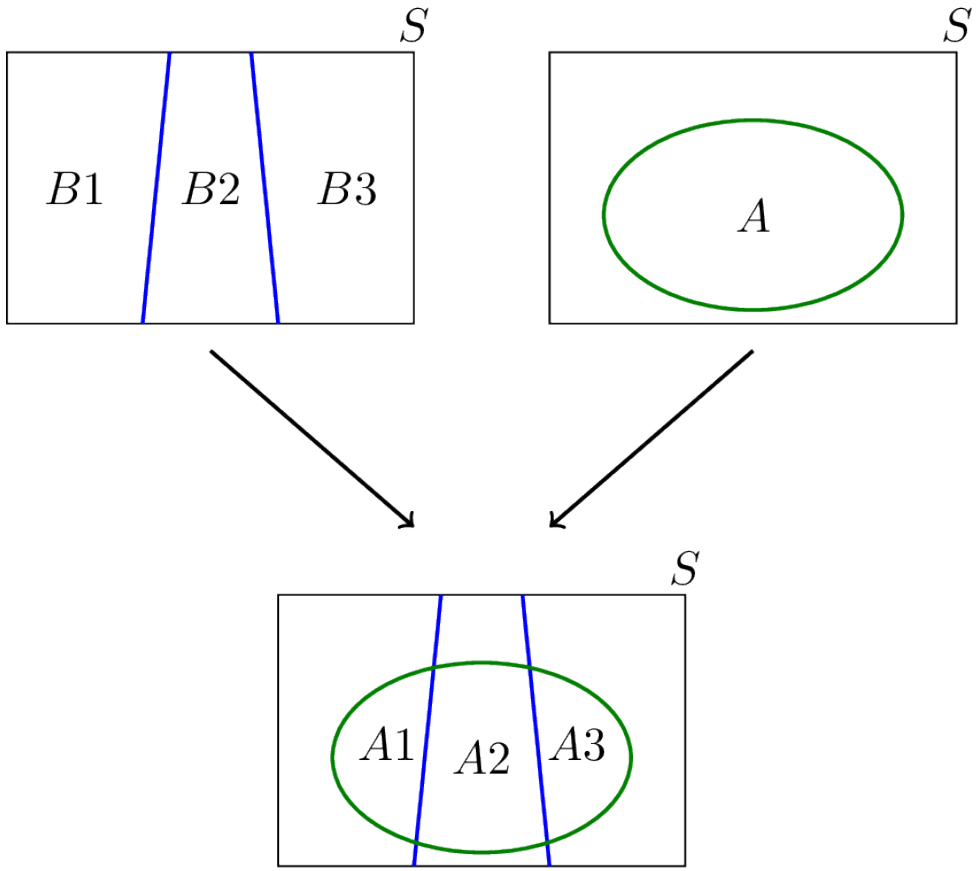
$$P(U|+) = \frac{P(U)P(+|U)}{P(+)} = \frac{P(U)P(+|U)}{P(U)P(+|U) + P(\bar{U})P(+|\bar{U})}$$

$$= \frac{0.005 \times 0.99}{0.005 \times 0.99 + 0.995 \times 0.001} \approx 33.2\%$$

※ $P(A) = \sum_{b \in B} P(b)P(A|b)$

※ Law of Total Probability

$$P(A) = \sum_{b \in B} P(b)P(A|b)$$



Naïve Bayes Classifier

- Conditional probability model for Naïve Bayes classifier
  - Naïve Bayes classifier calculates following probability for every class

$$p(C_k|x_1, \dots, x_p) = p(C_k|\mathbf{x})$$

    - x_i represents each feature (independent variable)
    - k represents k-th class and classifier assigns output class with the maximum probability

  - Re-formulation using Bayes' theorem

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

    - This equation is also written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Naïve Bayes Classifier

- Naïve Bayes Classifier

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

  - Denominator $p(\mathbf{x})$ does not depend on class

$$\underset{k}{\operatorname{argmax}}\, p(C_k|\mathbf{x}) = \underset{k}{\operatorname{argmax}}\, p(C_k)p(\mathbf{x}|C_k)$$

- $p(C_k)p(\mathbf{x}|C_k)$ is equivalent to the joint probability $p(C_k, x_1, \ldots, x_p)$
  - Using chain rule $p(C_k, x_1, \ldots, x_p)$ can be written as follows

$$\begin{aligned} p(C_k, x_1, \ldots, x_p) &= p(C_k)p(x_1, \ldots, x_p|C_k) = p(C_k)p(x_1|C_k)p(x_2, \ldots, x_p|C_k, x_1) \\ &= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1)\cdots p(x_p|C_k, x_1, \ldots, x_{p-1}) \end{aligned}$$

  - Naïve Bayes classifier assumes conditional independence of each feature

$$p(x_i|C_k, x_j) = p(x_i|C_k)$$
$$p(x_i|C_k, x_j, x_l) = p(x_i|C_k)$$

※ Chain Rule

- Chain rule permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities

$$P(A_n, \ldots, A_1) = P(A_n | A_{n-1}, \ldots, A_1) \cdot P(A_{n-1}, \ldots, A_1)$$

  - Repeating this process with each final term creates the product

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P\left(A_k \middle| \bigcap_{j=1}^{k-1} A_j\right)$$

Naïve Bayes Classifier

- Naïve Bayes Classifier

$$p(C_k)p(x_1|C_k)p(x_2|C_k,x_1)\cdots p\left(x_p\middle|C_k,x_1,\ldots,x_{p-1}\right)$$

$$= p(C_k)p(x_1|C_k)p(x_2|C_k)\cdots p\left(x_p\middle|C_k\right) = p(C_k)\prod_{i=1}^{p} p(x_i|C_k)$$

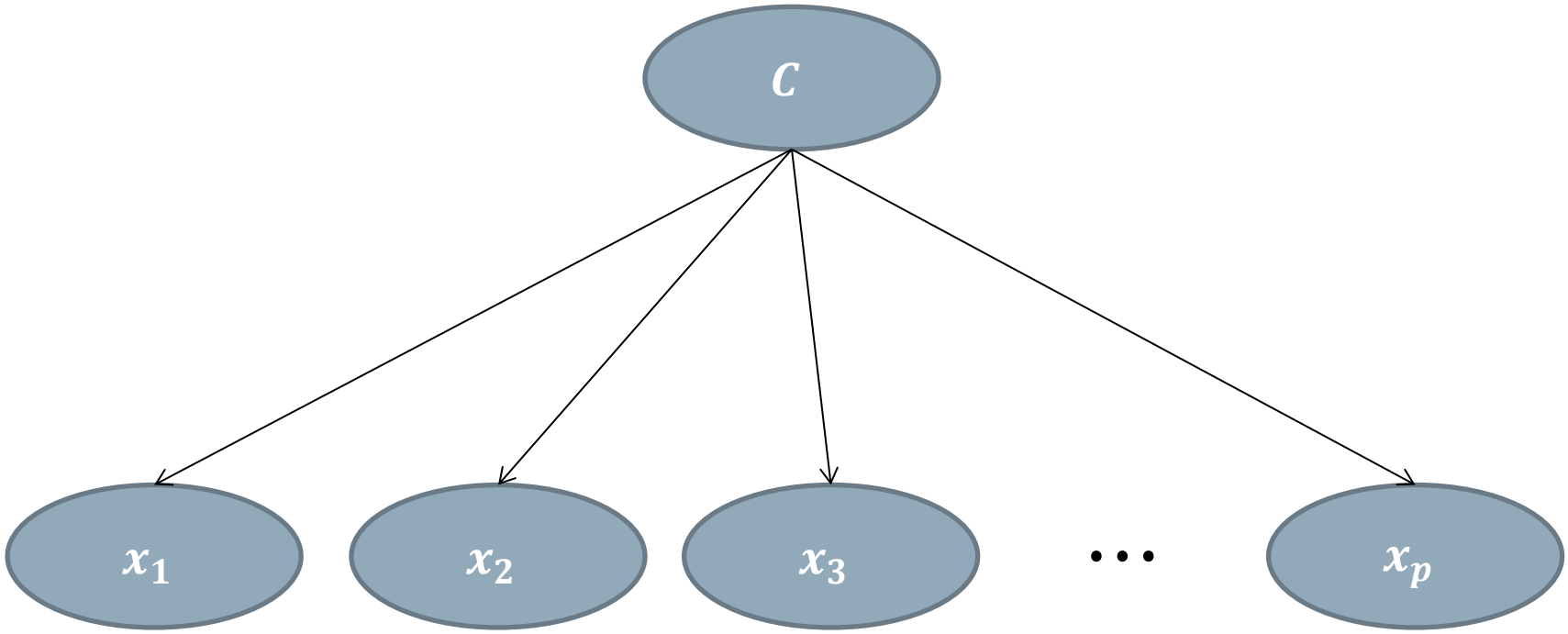
$$\therefore p(C_k)p(\mathbf{x}|C_k) = p(C_k)\prod_{i=1}^{p} p(x_i|C_k)$$

- Decision function of Naïve Bayes classifier

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}} p(C_k)\prod_{i=1}^{p} p(x_i|C_k)$$

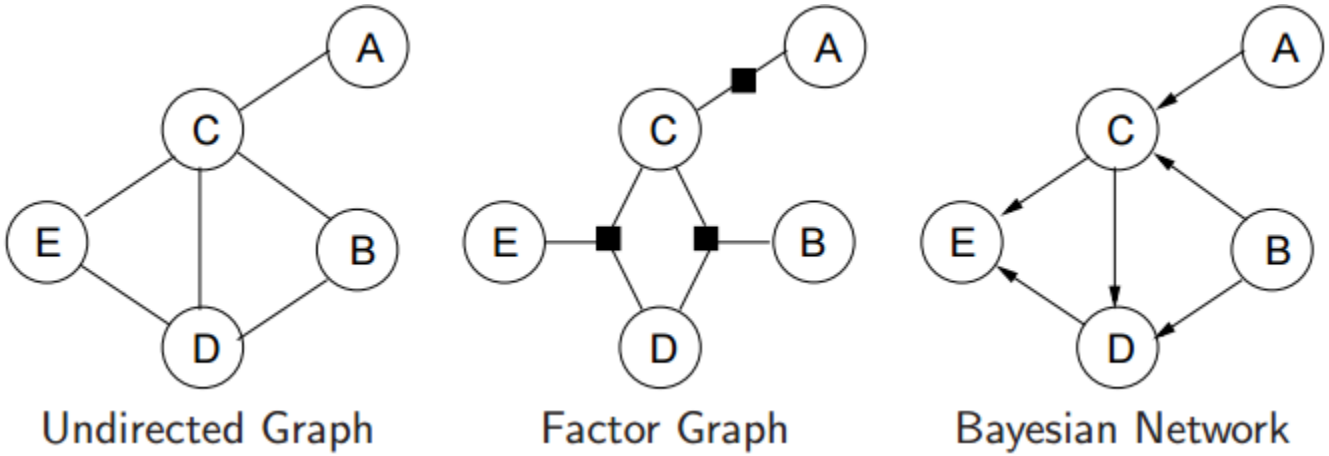
Naïve Bayes Classifier

- Graphical model representation of Naïve Bayes



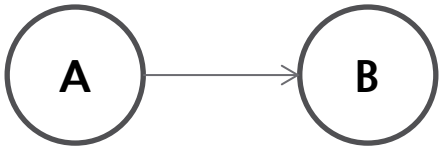
※ Graphical Model

- In graphical model, each node represents a random variable and the edge express probabilistic relationships between these variables



  - Bayesian network is a specific form of graphical model
    - Directed acyclic graphs
    - If two nodes A and B are connected by edge and direction of edge heads to B, it means that the state of A affects on probability of B

$$p(B, A) \neq p(B)p(A)$$
$$p(B|A) \neq p(B)$$



Naïve Bayes Classifier: Advantages and Limitation

- Advantages of Naïve Bayes
  - Simple and fast to train
  - Works well with small datasets
  - Handles missing data well
  - Requires less training data compared to other classifiers

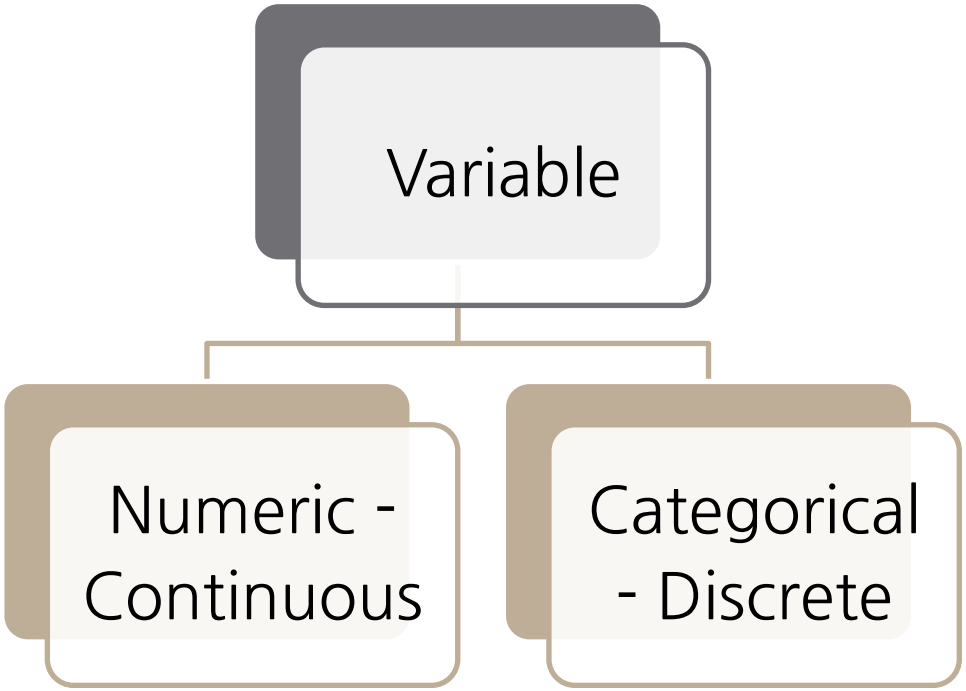
- Limitations of Naïve Bayes
  - Independence assumption is often unrealistic
  - Struggles with correlated features
  - Zero probability problem (solved using Laplace smoothing)

Naïve Bayes Classifier: Parameter Estimation

- Parameter estimation and event models
  - A class' prior setting: by assuming equiprobable classes ($p(C_k) = 1/K$) or by calculating an estimate for the class probability from the training set ($p(C_k) = n_k/n$)
  - Select appropriate probability distribution for $p(x_i|C_k)$
    - For continuous variables, Gaussian distribution is the common choice
    - For discrete variables, multinomial distribution is the common choice
  - After setting, probabilistic model for naïve Bayes classifier, parameters of distributions are estimated using training data
    - Calculate $\tilde{p}(C_{y_i}|\mathbf{x}) = (C_{y_j}) \prod_{i=1}^{p} p(x_i|C_{y_j})$ for j-th sample
    - Calculate $\tilde{p}(\boldsymbol{C}|\mathbf{X}) = \prod_{j=1}^{n} \tilde{p}(C_{y_i}|\mathbf{x})$ and maximize this probability

Naïve Bayes Classifier

- Determine $p(x_i|C_k)$
  - The probability functions depend on the type of variables
    - Probability distributions for discrete random variables: Binomial, Multinomial, Geometric, …
    - Probability distributions for continuous random variables: Gaussian(Normal), χ^2, beta, F, t, …



Bernoulli Naïve Bayes

- Bernoulli naïve Bayes
  - In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs
    - Ex) Each variable only takes values 0 or 1
  - Each x_i is a boolean expressing the occurrence of event

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{d} p_{ki}^{x_i} (1 - p_{ki})^{1-x_i}$$

    - d is the number of input features
    - p_{ki} is the probability that x_i is 1 (true) for class k

※ Bernoulli distribution

- The probability distribution of a random variable which takes the value 1 with success probability of p and the value 0 with failure probability of $q = 1 - p$
  - For random variable following Bernoulli distribution,

$$p(X = 1) = 1 - p(X = 0) = p = 1 - q$$

  - Probability mass function over possible outcomes y

$$f(y; p) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}$$

    - This can also be expressed as

$$f(y; p) = p^y (1 - p)^{1-y} \quad \text{for } y \in \{1,0\}$$

  - Expected value of a Bernoulli random variable X

$$\mathbb{E}[X] = p$$

  - Variance of a Bernoulli random variable

$$\text{Var}[X] = p(1 - p)$$

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model
  - Each input variable can take values 0 or 1
  - There exist n samples with d features
  - Prior probability $p(C_k)$

$$p(C_k) = \frac{n_k}{n}$$

    - n_k is the number of samples that y_i belongs to class k
  - Likelihood $p(\mathbf{x}|C_k)$

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{d} p_{ki}^{x_i} (1 - p_{ki})^{1-x_i}$$

  - Posterior probability $p(C_k|\mathbf{x})$

$$p(C_k|\mathbf{x}) \propto p(C_k)p(\mathbf{x}|C_k)$$

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model

$$L = \prod_{j=1}^{n} p\left(C_{y_j}\right) p\left(\mathbf{x}_j \middle| C_{y_j}\right) = \prod_{j=1}^{n} \frac{n_{y_j}}{n} \left(\prod_{i=1}^{d} p_{y_j i}^{x_{ji}} \left(1 - p_{y_j i}\right)^{1 - x_{ji}} \right)$$

$$\log L = \sum_{j=1}^{n} \log \frac{n_{y_j}}{n} + \sum_{j=1}^{n} \sum_{i=1}^{d} \left(x_{ji} \log p_{y_j i} + \left(1 - x_{ji}\right) \log\left(1 - p_{y_j i}\right) \right)$$

  - n is the total number of data samples
  - n_{y_j} is the number of data samples belong to class y_j $(y_j \in \{1,2, \ldots k\})$
  - x_{ji} is the i-th input variable's value for j-th data sample
- Parameters to be estimates
  - For each class k, probability to occur 1 for each feature i, p_{ki}

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model
  - To obtain optimal p_{ki}, set $\frac{\partial \log L}{\partial p_{ki}} = 0$

$$\frac{\partial \log L}{\partial p_{ki}} = \sum_{j \in \{m: y_m = k\}} \left\{ \frac{x_{ji}}{p_{ki}} - \frac{1 - x_{ji}}{1 - p_{ki}} \right\}$$

$$= \frac{n_{k1}}{p_{ki}} - \frac{n_{k0}}{1 - p_{ki}} = 0$$

  - $n_{k1} = |\{m: x_{mi} = 1, y_m = k\}|$ is the number of data samples in set of $\{m: x_{mi} = 1, y_m = k\}$
  - $n_{k0} = |\{m: x_{mi} = 0, y_m = k\}|$ is the number of data samples in set of $\{m: x_{mi} = 0, y_m = k\}$
  - $\{m: x_{mi} = 1, y_m = k\}$ is a set that contains every sample with $x_i = 1$ in class k
  - $\{m: x_{mi} = 0, y_m = k\}$ is a set that contains every sample with $x_i = 0$ in class k
  - $|\{m: x_{mi} = 1, y_m = k\}| + |\{m: x_{mi} = 0, y_m = k\}| = n_k$

$$p_{ki} = \frac{n_{ki}}{n_k}$$

Bernoulli Naïve Bayes: Estimation of Parameters

x	y
1	0
1	0
1	0
0	0
0	1
0	1
1	1

$$p(x = 0|y = 0) = p_{00} = \frac{1}{4}$$

$$p(x = 1|y = 0) = p_{01} = \frac{3}{4}$$

$$p(x = 0|y = 1) = p_{10} = \frac{2}{3}$$

$$p(x = 1|y = 1) = p_{11} = \frac{1}{3}$$

Categorical Naïve Bayes

- Discrete random variables with more than two outcomes

$$P\big(\mathbf{x} = (x_1, x_2, \ldots, x_m)\big) = \prod_{j=1}^{m} p_j^{x_j}$$

  - m outcomes
  - x_j is a binary indicator variable: 1 when j-th outcome is observed; otherwise 0
  - p_j is probability of j-th outcome $(\sum_{j=1}^{m} p_m = 1)$

$$p(\mathbf{X}|C_k) = \prod_{i=1}^{n} \prod_{j=1}^{m} p_{kj}^{x_j}$$

$$\log p(\mathbf{X}|C_k) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_j \log p_{kj}$$

$$p_{ki} = \frac{n_{kj}}{n_k}$$

Categorical Naïve Bayes

- Discrete random variables with more than two outcomes

x	y
High	0
Mid	0
High	0
Low	0
High	0
Low	0
High	0
High	0

$$p(x = High|y = 0) = p_{0,High} = \frac{5}{8}$$

$$p(x = Mid|y = 0) = p_{0,Mid} = \frac{1}{8}$$

$$p(x = Low|y = 0) = p_{0,Low} = \frac{2}{8}$$

5high → 6high $\frac{5+1}{8+3} = \frac{6}{11}$

1mid → 2mid $\frac{1+1}{8+3}$

2low → 3low $\frac{2+1}{8+3} = \frac{3}{11}$

Parameter Smoothing

- Smoothing techniques for parameter estimation
  - Smoothing techniques helps prevent zero probabilities when a category is not observed in the training data
  - Laplace smoothing (additive smoothing)

$$p_{kj} = \frac{n_{kj} + \alpha}{n_k + \alpha m}$$

    - α: smoothing parameter, commonly $\alpha = 1$

Multinomial Naïve Bayes

categorical이 하나의 범주값 (색상='빨강, 노랑, 파랑') 등이라면
multinomial은 횟수 빈도 (good이라는 단어 3번 등장)

- Multinomial naïve Bayes
  - With a multinomial event model, samples represent the frequencies with which certain events have been generated by a multinomial $(p_1, \ldots, p_d)$
    - p_i is the probability that event i occurs
    - m is the number of features in input data

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i}$$

    - $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ represents each sample and it can be seen as a histogram with x_i counting the number of times event i was observed in a particular instance
  - This is the event model typically used for document classification
    - With events representing the occurrence of a word in a single document → bag of words representation

※ Multinomial distribution

- Multinomial distribution is a generalization of the binomial distribution
  - Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments with success probability p

$$p(k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

  - Example of binomial distribution is the distribution of the number of head when flipping a coin n times (in this case, $p = 0.5$)
    - Probability that k times head occur among n trials

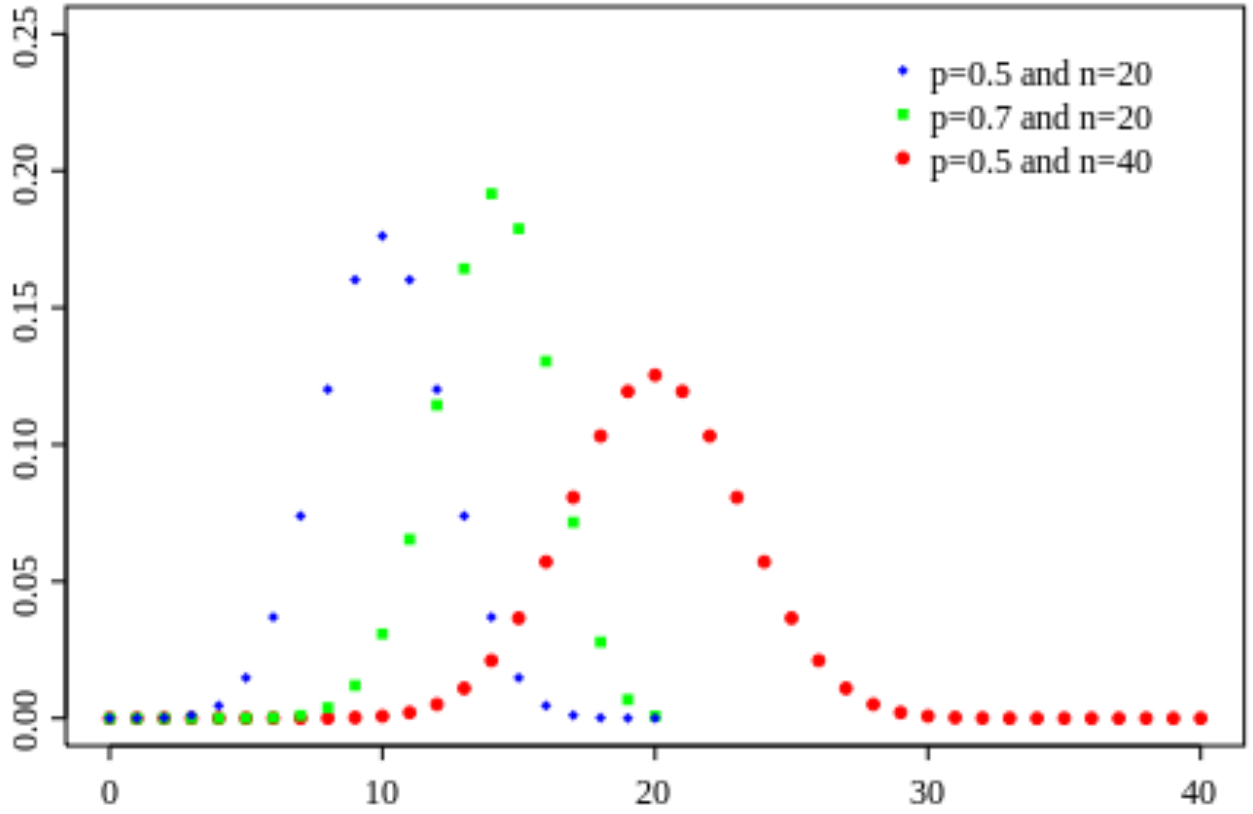
$$p(k) = \frac{n!}{k!\,(n-k)!} 0.5^k 0.5^{n-k} = \frac{n!}{k!\,(n-k)!} 0.5^n$$

  - In multinomial distribution, possible outcome is more than two and each outcome has its own probability to occur, $(p_1, \ldots, p_d)$
    - $p_1 + \cdots + p_d = 1$
    - d is the number of possible outcomes
    - $n_{\mathbf{x}} = \sum_{i=1}^{d} x_i$

$$p(\mathbf{x} = (x_1, x_2, \ldots, x_d)) = \frac{n_{\mathbf{x}}!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d}$$

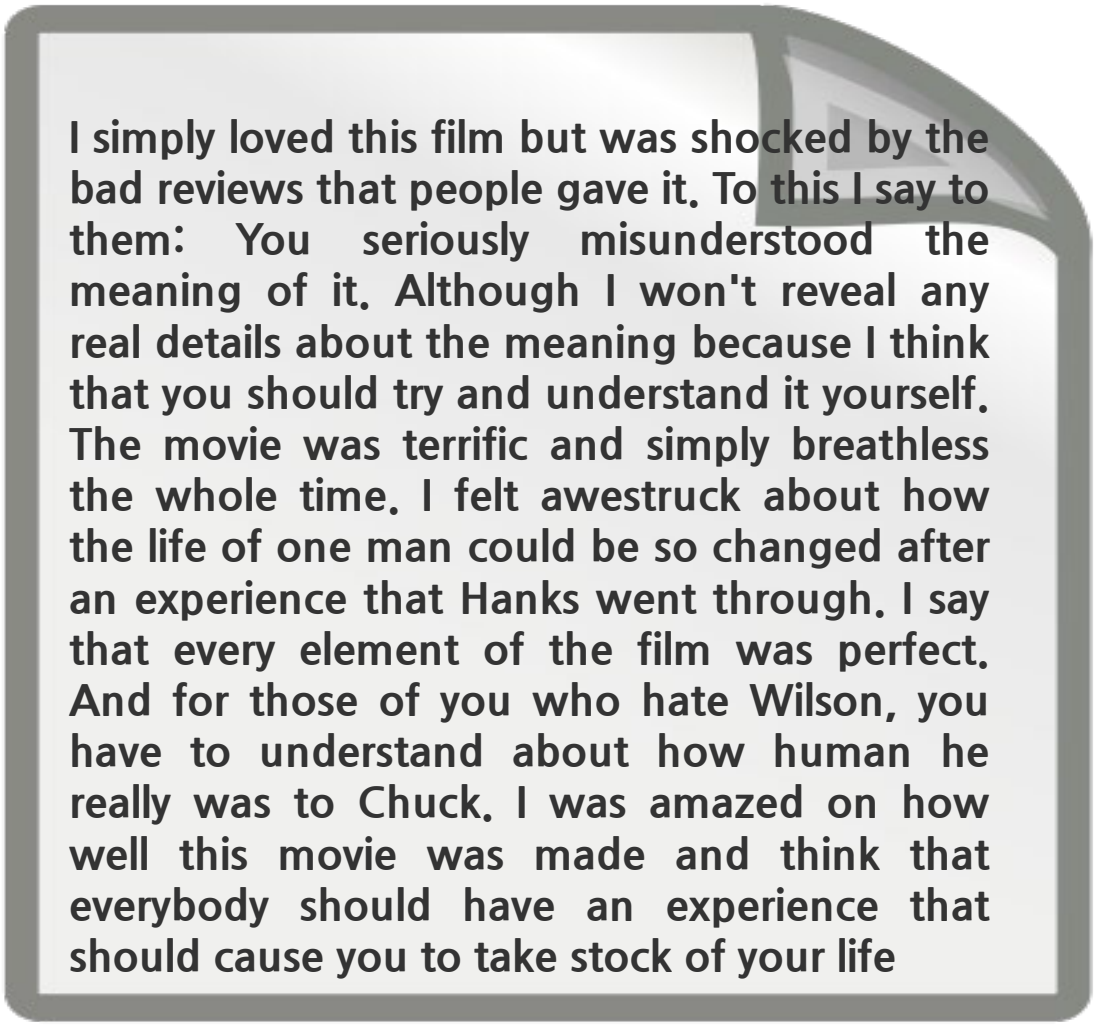
※ Multinomial distribution

- Binomial distribution



Multinomial Naïve Bayes: Application

- Relies on very simple representation of document
  - Bag of words

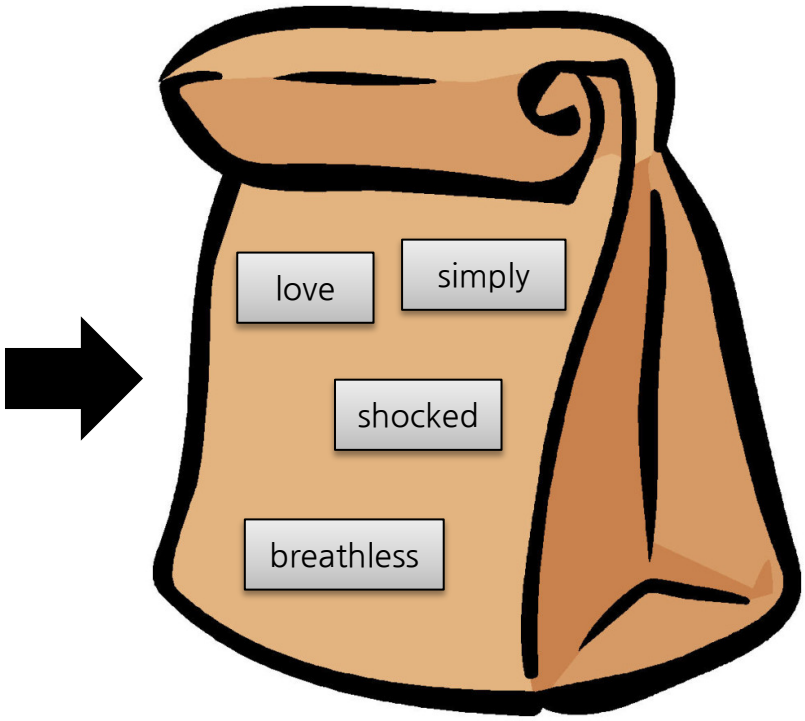


$$f(\quad) = C$$

Multinomial Naïve Bayes: Application

- The bag of words representation

I simply loved this film but was shocked by the bad reviews that people gave it. To this I say to them: You seriously misunderstood the meaning of it. Although I won't reveal any real details about the meaning because I think that you should try and understand it yourself. The movie was terrific and simply breathless the whole time. I felt awestruck about how the life of one man could be so changed after an experience that Hanks went through. I say that every element of the film was perfect. And for those of you who hate Wilson, you have to understand about how human he really was to Chuck. I was amazed on how well this movie was made and think that everybody should have an experience that should cause you to take stock of your life



Multinomial Naïve Bayes: Application

- The bag of words representation
  - It is possible to select a subset of words

$$f($$

Word	Frequency
love	1
great	2
recommend	1
simply	1
happy	1
bad	2
⋮	⋮

$$) = C$$

Multinomial Naïve Bayes: Application

- The bag of words representation
  - It is possible to select a subset of words

Word	Frequency
love	1
great	2
recommend	1
old	0
simply	1
happy	1
bad	2
dog	0
⋮	⋮

$$f(\quad) = C$$

Multinomial Naïve Bayes: Application

- Revisit naïve Bayes

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{p} p(x_i | C_k)$$

  - Each ith column represent the specific word
  - C_k is the possible output class
    - Ex.) Spam filter: spam or non-spam

- The basic idea that uses Naïve Bayes for text classification
  - Each class has the different distribution of words
    - Spam mails usually contain “click below”, “free dvd”, “great offer” and etc.

Multinomial Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimator with Laplace smoothing

$$p_{ki} = \frac{\sum_{j \in C_k} x_{ji} + \alpha}{\sum_{j \in C_k} \sum_{l=1}^{d} x_{jl} + \alpha d}$$

  - p_{ki}: The probability of occurrence of i-th variable for class k
  - d: The number of input variables
  - x_{ji}: The occurrence of i-th variable for sample j

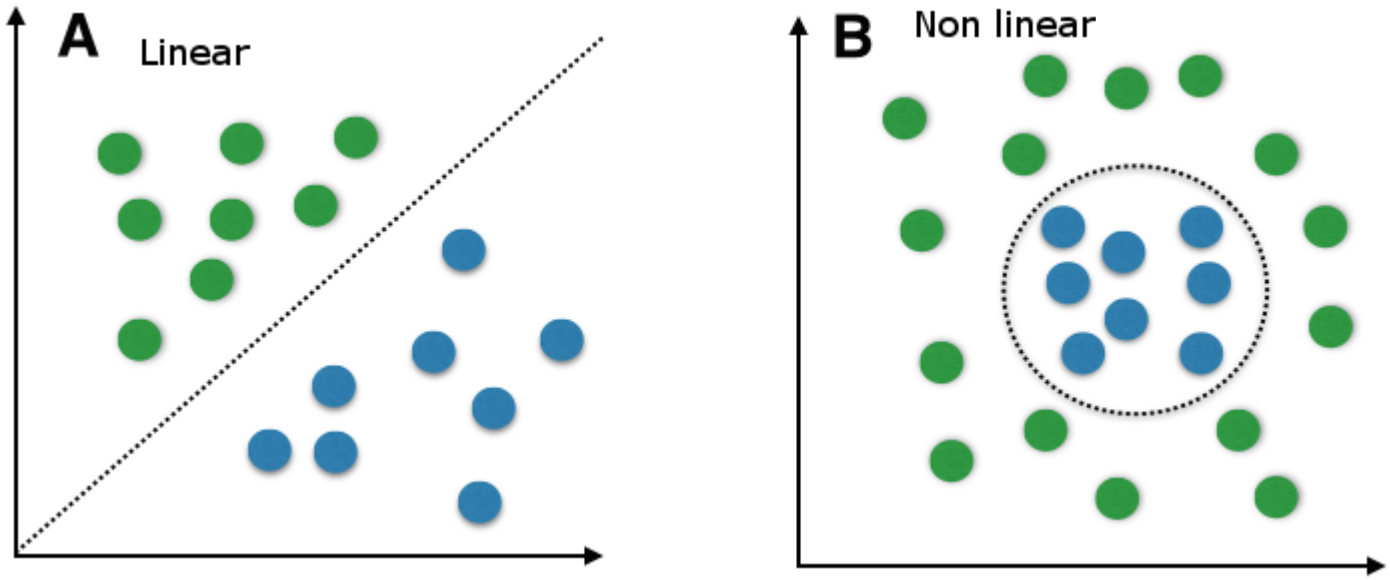
Multinomial Naïve Bayes

- Multinomial naïve Bayes
  - The multinomial naive Bayes classifier becomes a linear classifier when expressed in log-space

$$\log p(C_k|\mathbf{x}) \propto \log p(C_k) \prod_i p_{k_i}^{x_i} = \log p(C_k) + \sum_{i=1}^{n} x_i \log p_{ki} = b + \mathbf{w}_k^T \mathbf{x}$$

    - $b = \log p(C_k)$
    - $w_{ki} = \log p_{ki}$
    - $\frac{(\sum_i x_i)!}{\prod_i x_i!}$ term only depends on $\mathbf{x}$ and does not depend on class

※ Decision Boundary



Complement Naïve Bayes

- Complement naïve Bayes
  - Complement naïve Bayes is an adaptation of the standard multinomial naïve Bayes algorithm that is particularly suited for imbalanced data sets
  - Unlike multinomial naïve Bayes, it uses word occurrence frequencies from the complement of each class

$$p_{ki} = \frac{\sum_{j \notin C_k} x_{ji} + \alpha}{\sum_{j \notin C_k} \sum_{l=1}^{d} x_{jl} + \alpha d}$$

Gaussian Naïve Bayes

- Gaussian naïve Bayes
  - When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-u_k)^2}{2\sigma_k^2}}$$

$$p(C_k)\prod_{i=1}^{p} p(x_i|C_k) = p(C_k)\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{(v_i-u_{ki})^2}{2\sigma_{ki}^2}}$$

Naïve Bayes Classifier: Variations

- Semi-naïve Bayes classifier
  - A type of Bayesian classifier that relaxes the strict independence assumption of the Naive Bayes classifier, allowing for limited dependencies between features, while still maintaining computational efficiency and simplicity
    - It assumes that correlations exist only within disjoint subsets of features, meaning that features within a subset can depend on each other, but not across subsets

$$P(\mathbf{x}|C) = P(x_1, x_2, \ldots, x_n|C)$$
$$= P(x_1|C)P(x_2|C) \cdot \cdots \cdot P(x_k|C) \cdot P(x_{k+1}, x_{k+2}, \ldots, x_n|C)$$

    - $x_1, x_2, \ldots, x_k$ are the features assumed to be conditionally independent
    - $x_{k+1}, x_{k+2}, \ldots, x_n$ are the features that are allowed to have dependencies

Naïve Bayes Classifier: Variations

- Hidden naïve Bayes classifier
  - The Hidden naïve Bayes classifier is an extension of the standard naïve Bayes classifier, which introduces the concept of latent or hidden variables into the model
    - Hidden variables are assumed to influence the observed features but are not directly observed or measured

$$P(C|\mathbf{x}) = \sum_{h \in H} P(C|h) \prod_{i} P(x_i|C,h)$$

    where h represents the hidden variables
  - This model is especially useful when we suspect that there are unobserved or hidden factors influencing the outcome, and these factors should be taken into account for better classification performance

Naïve Bayes Classifier: Variations

- Bayesian network classifier
  - A Bayesian network classifier a probabilistic classifier that uses a **Bayesian network** to model the relationships between variables
  - It allows dependencies between features by using a probabilistic graphical model

$$P(\mathbf{x}) = \prod_{i=1} P(x_i|pa_i)$$

    - pa_i represents the set of parent nodes for x_i
  - It requires structure learning which can be computationally expensive, especially for large datasets with many features