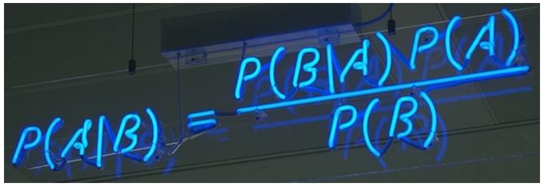
DATA SCIENCE

COMP2200/6200

09 – Naïve Bayes Classifier



Xuyun Zhang (School of Computing)

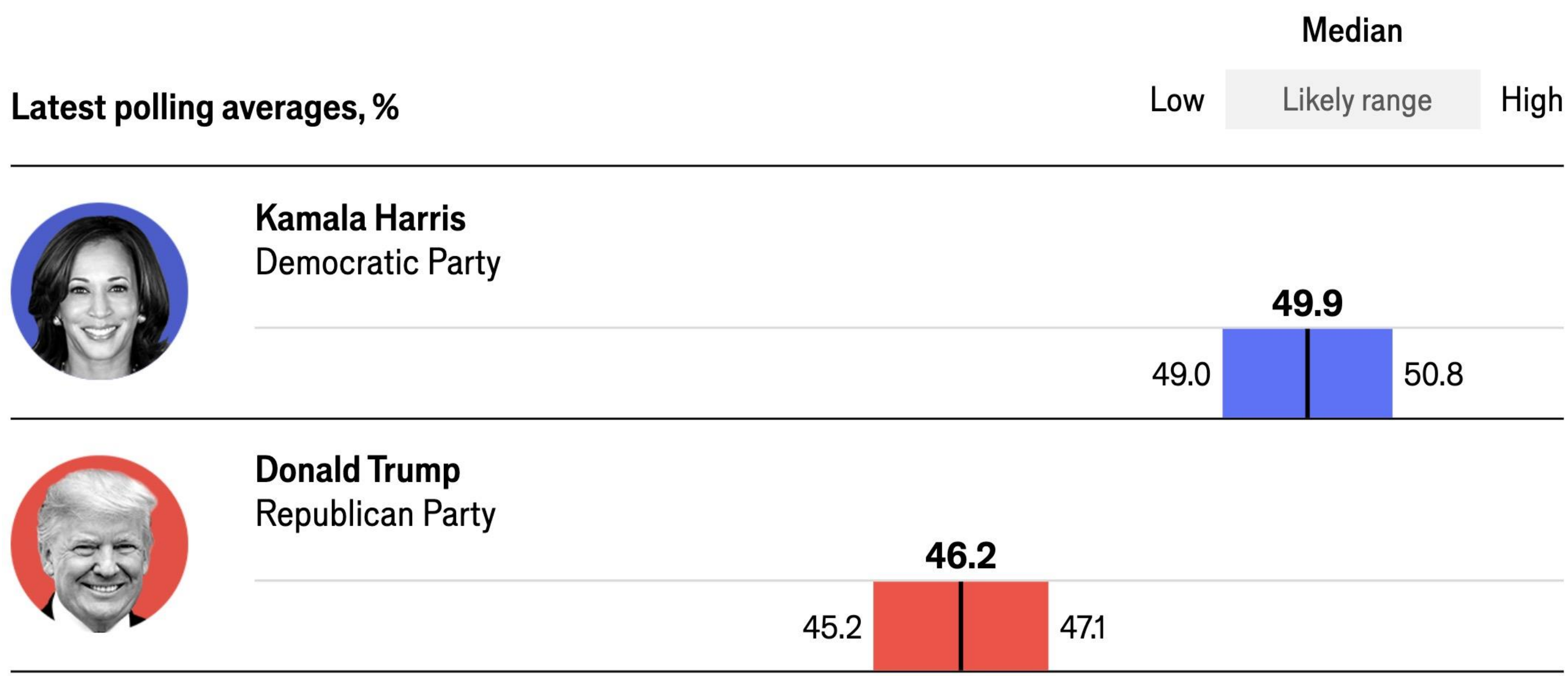
Lecture Outline

- Probability Basics
- Naïve Bayes Classifier
- Practical

Loss of Certainty

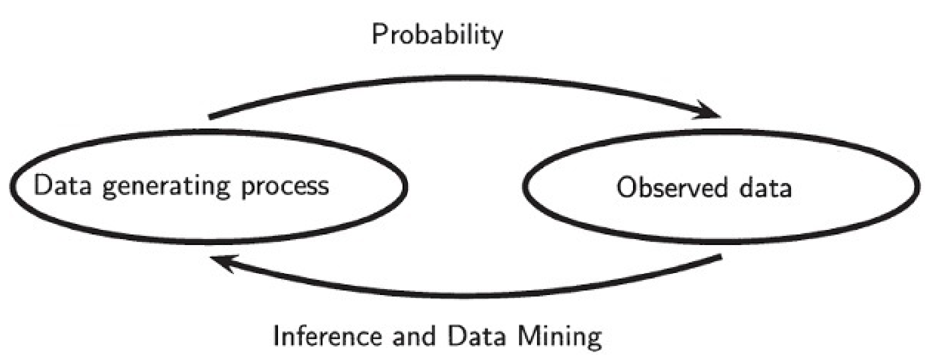
MACQUARIE University

Last updated on October 1st 2024



Loss of Certainty (Cont'd)

- Uncertainty
  - Limited knowledge in partially observable environments
    - E.g., measurements in meteorology, physics, and engineering
  - Feature outcome in stochastic environments
    - E.g., who will be the next US president?
- Probability theory and statistics
  - Consistent framework quantifying uncertainty
  - Central foundation for data science and machine learning



Probability

- Basic concepts
  - Random experiment, a.k.a., trial
    - E.g., flipping a coin, or tossing a die
  - Outcome: result of a single experiment execution
  - Sample space Ω: the set of all possible outcomes
    - E.g., {H, T} for coin flipping
    - E.g., {1, 2, 3, 4, 5, 6} for die tossing
  - Probability distribution: a function provide the probabilities of occurrence of different possible outcomes in an experiment
  - Event: a set contains 0 or more outcomes, i.e., a subset of Ω
    - E.g., {2, 4, 6}
  - Event space S: the power set of Ω, i.e., $S = 2^S$
  - Probability measure function $\Pr: S \rightarrow \mathbb{R}$

Discrete Random Variables

- Random variable X can take a discrete number of values from a set: $\{x_1, x_2, \dots, x_n\}$
- $p(x) \equiv \Pr\{X = x\}$ is the probability that X takes the specific value x
- $p(x)$ is the probability mass function
  - To be different from the continuous case
- Non-negative $0 \leq p(x) \leq 1$
- Total probability

$$\sum_x p(x) = 1$$

How to Get Probability?

- If we assume that each outcome occurs with an equal chance, the probability is $\frac{1}{|\Omega|}$ (theoretical probability)
  - E.g., fair coin flipping, $\Pr\{X = H\} = \Pr\{X = T\} = 1/2$
  - E.g., fair die tossing, $\Pr\{X = 1\} = \cdots = \Pr\{X = 6\} = 1/6$
- Estimation from relative frequency
  - Repeat random experiments N times
  - Count the occurrence/frequency of an outcome x, N_x
  - Relative frequency $\frac{N_x}{N}$
  - Estimate $\Pr\{X = x\} = \lim_{N \to +\infty} \frac{N_x}{N}$ (law of large numbers)
  - In practice, make N large enough to have a good estimation

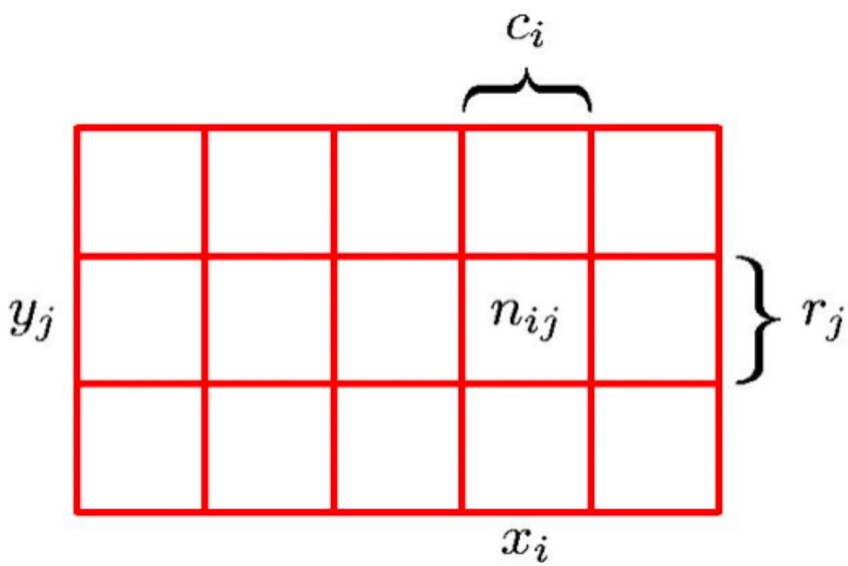
Random Variable Relationships

MACQUARIE University

- What's really interesting is the relationships among random variables, e.g., Height vs Weight
- Joint probability of two random variables X and Y
  - $p(x, y) \equiv \Pr\{X = x, Y = y\}$
  - X and Y are independent $\Leftrightarrow p(x, y) = p(x)p(y)$
- Conditional probability of $X = x$ given $Y = y$
  - $p(x|y) = \frac{p(x,y)}{p(y)} \quad \Leftrightarrow \quad p(x, y) = p(y)p(x|y)$ (product rule)
  - If X and Y are independent, $p(x|y) = ?$
- Marginal probability
  - $p(x) = \sum_y p(x, y) = \sum_y p(y)p(x|y)$ (sum rule)

Examples

MACQUARIE University



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

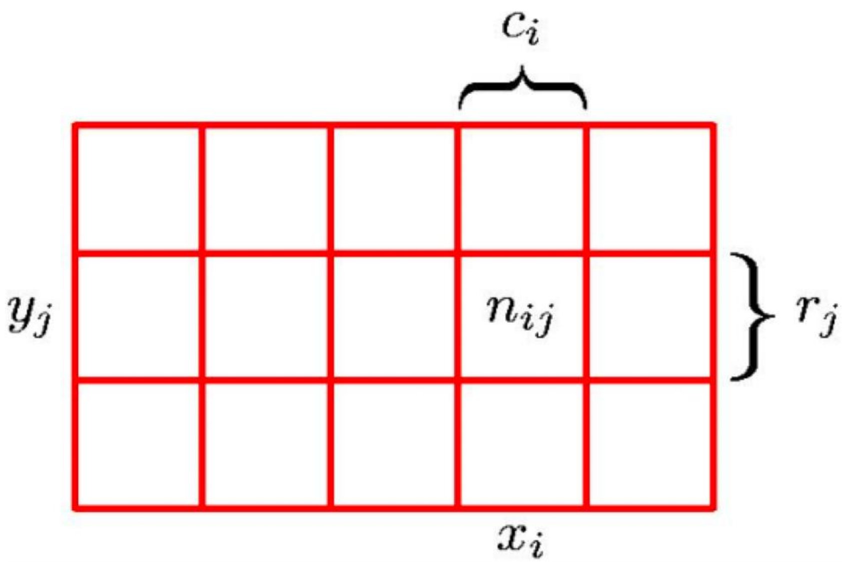
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Examples (Cont'd)



Sum Rule

$$
\begin{aligned}
p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij} \\
&= \sum_{j=1}^{L} p(X = x_i, Y = y_j)
\end{aligned}
$$

Product Rule

$$
\begin{aligned}
p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\
&= p(Y = y_j | X = x_i) p(X = x_i)
\end{aligned}
$$

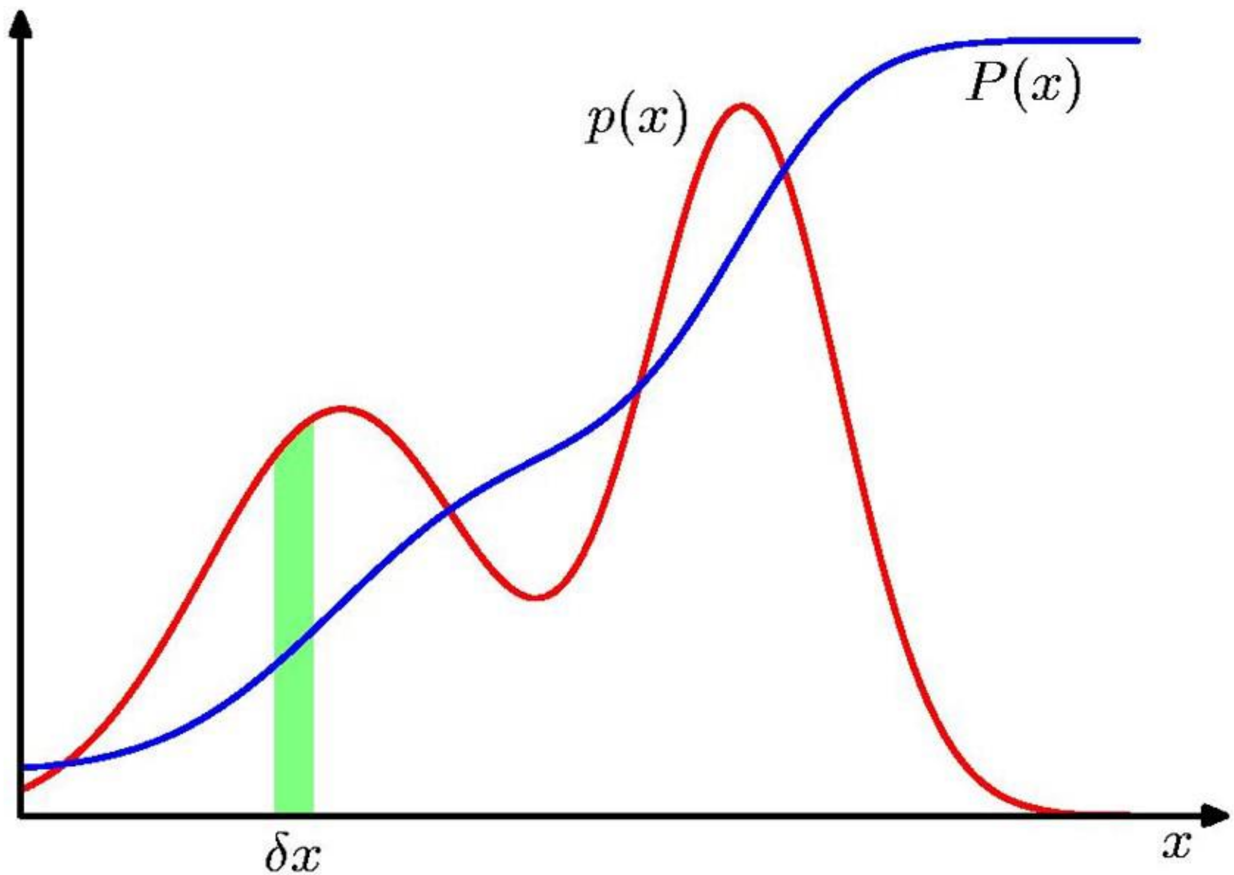
Continuous Random Variables

MACQUARIE University

- X can take values on a continuous range, i.e., $S \subset \mathbb{R}$
  - E.g., height, weight
  - $\Pr\{\mathrm{X} = x\} = ?$
  - Probability of X falling into $(x, x + \delta x)$: $\Pr\{\mathrm{X} \in (x, x + \delta x)\}$
- Probability density function (PDF) $p(x)$ over x
  - $p(x) = \lim_{\delta x \to 0} \Pr\{\mathrm{X} \in (x, x + \delta x)\}$
  - Probability of X in (a, b): $\Pr\{\mathrm{X} \in (a, b)\} = \int_a^b p(x)\mathrm{d}x$
- Cumulative distribution function (CDF) $P(x)$ over x

$$P(x) = \Pr\{X \leq x\} = \Pr\{\mathrm{X} \in (-\infty, x)\} = \int_{-\infty}^{x} p(t)\mathrm{d}t$$

PDF and CDF



Distribution Properties

- Non-negative

$$p(x) \geq 0$$

- Normalized

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

  - Is it possible $p(x) \geq 1$?

- Relationship between PDF and CDF

$$p(x) = P'(x) = \frac{\mathrm{d}P(x)}{\mathrm{d}x}$$

- Extension to random vector (multivariate distribution)

Random Variable Relationship

- Joint density of two random variable: $p(x, y)$
- Conditional density of X given $Y = y$

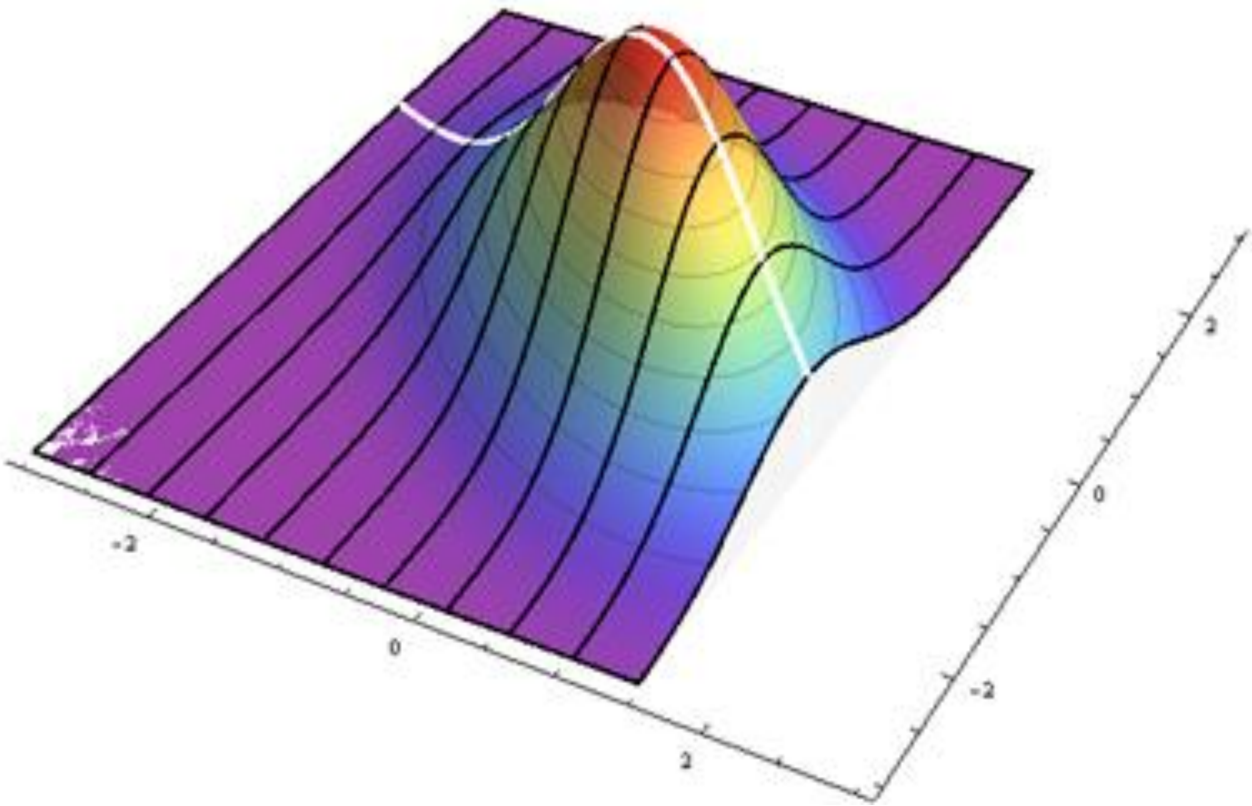
$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- Marginal density

$$p(x) = \int_y p(x, y) = \int_y p(x|y)p(y)$$

  - Sum rule
  - Product rule

Examples



Independence

- Random variables X and Y are independent:
  - Can be denoted as $X \perp Y$
  - $(X \perp Y) \Longrightarrow P(X, Y) = P(X)P(Y)$
  - $(X \perp Y) \Longleftarrow P(X \mid Y) = P(X)$ or $P(Y) = 0$

- Conditional independence
  - $(X \perp Y \mid Z) \Longrightarrow P(X, Y \mid Z) = P(X|Z)P(Y|Z)$
  - $(X \perp Y \mid Z) \Longleftarrow P(X \mid Y, Z) = P(X|Z)$ or $P(Y, Z) = 0$

Expectation

MACQUARIE University

- X is a random variable, its expectation $\mathbb{E}[X]$

$$\mathbb{E}[X] \equiv \sum_x x \cdot p(x)$$

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x \cdot p(x) \mathrm{d}x$$

  - Expected outcome of a fair die ?
    - $(1+2+3+4+5+6) \cdot \frac{1}{6} = 3.5$

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + \mathrm{b}$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $(X \perp Y) \Longrightarrow \mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Variance

- Variance measures the deviation of X from its mean

$$\mathrm{Var}[X] = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2]$$

$$= \mathbb{E}[X^2] - \mathbb{E}^2[X]$$

- Standard deviation

$$\sigma[\mathrm{X}] = \sqrt{\mathrm{Var}[X]}$$

- $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$
- $(X \perp Y) \Longrightarrow \mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Lecture Outline

MACQUARIE University

- Probability Basics

- Naïve Bayes Classifier

Probabilistic Modelling

- How can we model classification with probability?
- Each attribute is a random variable
  - Target is a discrete random variable with distribution $P(\mathcal{C})$
  - Features are a random vector with joint distribution $P(X)$
  - Whole data can be captured by joint distribution $P(X, \mathcal{C})$
- Then, classification problem is to estimate $P(\mathcal{C}|X = \boldsymbol{x})$
  - $\boldsymbol{x}$ is the feature vector of a testing instance
  - Distribution $P(\mathcal{C}|\boldsymbol{x})$ can tell the probability of each class label
  - The label with the highest probability is the predicted label

$$\text{Label}(x) \leftarrow \arg\max_{\mathcal{C}_k}\{p(\mathcal{C}_k|\boldsymbol{x})\}$$

    - This can minimize classification error

Probabilistic Modelling (Cont'd)

MACQUARIE University

- The key is to estimate conditional distribution $P(\mathcal{C}|X)$
  - Question: is $\mathcal{C}$ and X independent ?
- In terms of product rule

$$P(\mathcal{C}|X) = \frac{P(\mathcal{C},\ X)}{P(X)}$$

- Option 1: estimate joint distribution $P(\mathcal{C},\ X)$ directly
  - Then, $P(X)$ can be estimate by sum rule

$$P(X) = \sum\nolimits_{\mathcal{C}} P(\mathcal{C}, X)$$

  - This is Bayes optimal classifier

Bayes Optimal Classifier

- Bayes' theorem (product rule)

$$P(C|x) = \overbrace{\frac{\overbrace{P(C)}^{\textbf{prior}}\quad\overbrace{P(x|C)}^{\textbf{class-conditional probability} \text{ OR } \textbf{likelihood}}}{\underbrace{P(x)}_{\textbf{evidence}}}}^{\textbf{posterior}}$$

- Bayes optimal classifier with minimum classification error: MAP (Maximum A Posterior) decision rule

$$Label(\boldsymbol{x}) \leftarrow \arg\max_{C_k} P(C_k|\boldsymbol{x})$$

Example

- Data

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example (Cont'd)

MACQUARIE University

- Joint distribution $P(\mathcal{C}, X)$ estimated from data

	X_1	X_2	X_3	X_4	$\mathcal{C}$	$P(\mathcal{C}, X)$
V_1	Sunny	Hot	High	Strong	Yes	0/14
V_2	Sunny	Hot	High	Weak	Yes	0/14
V_3	Sunny	Hot	Normal	Strong	Yes	0/14
V_4	Sunny	Hot	Normal	Weak	yes	0/14
	…	…	…	…	…	
V_{71}	Rain	Cool	Normal	Strong	No	1/14
V_{72}	Rain	Cool	Normal	Weak	No	0/14

- Why 72 possible values ?
  - $3 * 3 * 2 * 2 * 2 = 72$
- Many zero probabilities !

Example (Cont'd)

- Marginal distribution $P(X)$ estimated from data

	X_1	X_2	X_3	X_4	C	$P(X)$
V_1	Sunny	Hot	High	Strong	Yes/no	1/14
V_2	Sunny	Hot	High	Weak	Yes/no	1/14
V_3	Sunny	Hot	Normal	Strong	Yes/no	0/14
V_4	Sunny	Hot	Normal	Weak	Yes/no	0/14
	…	…	…	…	…	
V_{35}	Rain	Cool	Normal	Strong	Yes/no	1/14
V_{36}	Rain	Cool	Normal	Weak	Yes/no	1/14

- Sum rule: $P(X) = P(yes, X) + P(no, X)$

Example (Cont'd)

- Predicting a test instance $\boldsymbol{x}_1$(Day 15) and $\boldsymbol{x}_2$(Day 16)

Day	X_1	X_2	X_3	X_4	C
15	Sunny	Cool	High	Strong	?
16	Sunny	Hot	High	Strong	?

- $p(yes|\boldsymbol{x}_1) = \frac{p(yes,\boldsymbol{x}_1)}{p(\boldsymbol{x}_1)} = \frac{0}{0}$; $p(no|\boldsymbol{x}_1) = \frac{p(no,\boldsymbol{x}_1)}{p(\boldsymbol{x}_1)} = \frac{0}{0}$
  - Can't do an appropriate estimate due to lack of data!
- $p(yes|\boldsymbol{x}_2) = \frac{p(yes,\boldsymbol{x}_2)}{p(\boldsymbol{x}_2)} = \frac{0}{1} = 0$
- $p(no|\boldsymbol{x}_2) = \frac{p(no,\boldsymbol{x}_2)}{p(\boldsymbol{x}_2)} = \frac{1}{1} = 1 > p(yes|\boldsymbol{x}_2) \Longrightarrow C = no$

Observations

- Observations
  - Many parameters (joint probability) need estimation
  - $P(X)$ is constant w.r.t. C

$$P(C|X) \propto P(C)P(X|C)$$

- Maximizing $P(C|X) \Leftrightarrow$ maximizing $P(C)P(X|C)$

$$Label(\boldsymbol{x}) \leftarrow \arg\max_{C_k} P(C_k)P(\boldsymbol{x}|C_k)$$

- Now the problem is to estimate $P(C)$ and $P(X|C)$
  - Assumption: probabilities follows known distributions
  - E.g., multinomial distribution or Gaussian distribution
  - Then, the distribution parameters will be estimated

Naïve Bayes Classifier

MACQUARIE University

- Challenge of estimating $P(X|C)$: still need to model the joint probability of all features of $\boldsymbol{x} = \langle x_1, x_2, \dots, x_M \rangle$
  - Many parameters to estimate (much more than # of features)
- Naïve Bayes assumption: all input features are conditionally independent of each other
  - Strong assumption (hard to achieve in reality)

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{C}_k) &= p(x_1, \cdots, x_M|\mathcal{C}_k) \\
&= p(x_1|x_2, \cdots, x_M, \mathcal{C}_k) p(x_2, \cdots, x_M|\mathcal{C}_k) \\
&= p(x_1|\mathcal{C}_k) p(x_2, \cdots, x_M|\mathcal{C}_k) \\
&= p(x_1|\mathcal{C}_k) \cdots p(x_M|\mathcal{C}_k) = \prod_{i=1}^{M} p(x_i|\mathcal{C}_k)
\end{aligned}
$$

Product of individual probabilities

Naïve Bayes Classifier (Cont'd)

MACQUARIE University

Classification is easy, just have probabilities multiplied

- Now comes the decision rule

$$Label(\mathbf{x}) \leftarrow \arg\max_{C_k} P(C_k) \prod_{i=1}^{M} p(x_i|C_k)$$

- Scalable (linear # of parameters w.r.t. M features)
- In practice, we just need to simply compute the relative frequencies from training data to estimate the probabilities $P(C_k)$ and $p(x_i|C_k)$
- Performance is comparably good even though the conditionally-independent assumption is very strong

Distribution Examples

- Gaussian: Gaussian class-conditional distribution

$$p(\boldsymbol{x}|\mathcal{C}_k) = \prod_{i=1}^{M} \mathcal{N}(x_i|\mu_{ki}, \sigma_{ki}^2)$$

- Bernoulli: binary input features

$$p(\boldsymbol{x}|\mathcal{C}_k) = \prod_{i=1}^{M} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

- # of model parameters is significantly reduced

Probability Estimation

- Training is to estimate probabilities / parameters
  - Use maximum likelihood estimation for model parameters
- Estimate $p(C_k)$ from data set D
  - D_{C_k}: set of instances with C_k as class label

$$p(C_k) = \frac{|\mathcal{D}_{C_k}|}{|\mathcal{D}|}$$

- Estimate $p(x_i|C_k)$ for discrete feature x_i
  - D_{C_k, x_i}: set of instances with C_k as the class label and x_i as the value of its i^{th} feature

$$p(x_i|C_k) = \frac{|\mathcal{D}_{C_k,x_i}|}{|\mathcal{D}_{C_k}|}$$

Learning is easy, just create probability tables.

Probability Estimation (Cont'd)

- Estimate $p(x_i|C_k)$ for continuous feature x_i
  - Need to assume a distribution for the continuous feature
- Example: Gaussian distribution

$$p(x_i|\mathcal{C}_k) \sim \mathcal{N}(\mu_{\mathcal{C}_k,i}, \sigma^2_{\mathcal{C}_k,i})$$

  - $\mu_{C_k,\,i}$: mean of the i^{th} feature value of the instances with C_k as the class label
  - $\sigma^2_{C_k,i}$: the corresponding variance

$$p(x_i|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{C}_k,i}} \exp\left(-\frac{(x_i - \mu_{\mathcal{C}_k,i})^2}{2\sigma^2_{\mathcal{C}_k,i}}\right)$$

Example: Training/Learning

MACQUARIE University

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

P(Play=*No*) = 5/14

P(Play=*Yes*) = 9/14

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temp.	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Example: Testing

- Predict a new data instance
  - **x**′=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
- Calculate lookup tables

P(Outlook=*Sunny*|Play=*Yes*) = 2/9
P(Temperature=*Cool*|Play=*Yes*) = 3/9
P(Huminity=*High*|Play=*Yes*) = 3/9
P(Wind=*Strong*|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*No*) = 5/14

- Apply MAP rules

P(*Yes*|**x**′): [P(*Sunny*|*Yes*)P(*Cool*|*Yes*)P(*High*|*Yes*)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053
P(*No*|**x**′): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

  - **P(*Yes*|x′) < P(*No*|x′) → x′ to be labelled as "*No*"**

Example (Cont'd)

MACQUARIE University

- Continuous-valued features, e.g., Temperature

Day	D_1	D_2	…	D_{13}	D_{14}
Temperature	27.3	Cool	…	19.8	15.1
PlayTennis	No	No	…	Yes	No

  - Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8
  - No: 27.3, 30.1, 17.4, 29.5, 15.1

- Assume Temperature follows Gaussian distribution

- Estimate mean and variance for each class
  - $\mu_{Yes,\, temp} = 21.64,\ \sigma^2_{Yes,temp} = 2.35^2$
  - $\mu_{No,\, temp} = 23.88,\ \sigma^2_{No,temp} = 7.09^2$

Example (Cont'd)

- **Training stage**: output two Gaussian models with the above model parameters for two classes, respectively

$$P(x|Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2 \times 2.35^2}\right)$$
$$= \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$P(x|No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2 \times 7.09^2}\right)$$
$$= \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

- **Testing stage**: use the models to calculate probabilities

$$P(Temp = 25.0|Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(25.0-21.64)^2}{11.09}\right) = 0.06134$$

Zero conditional probability

MACQUARIE University

- If a combination of a class label and a (discrete) feature value is missing in training data, zero-conditional probability occurs
  - Possible (and reasonable) when training data set is small
  - Resultant conditional probability will always be zero, as product of the probabilities is computed during testing
  - Information provided by other features will be suppressed
- Remedy: smoothing, e.g., **Laplacian correction**

$$p(x_i|\mathcal{C}_k) = \frac{|\mathcal{D}_{\mathcal{C}_k,x_i}| + 1}{|\mathcal{D}_{\mathcal{C}_k}| + \gamma_i}$$

  - γ_i: # of possible feature values

Example:

- Predict a new data instance
  - $\mathbf{x}'$=(Outlook=*Overcast*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
- Calculate lookup tables

P(Outlook=*Overcast*|Play=*Yes*) = 4/9
P(Temperature=*Cool*|Play=*Yes*) = 3/9
P(Huminity=*High*|Play=*Yes*) = 3/9
P(Wind=*Strong*|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

P(Outlook=*Overcast*|Play=*No*) = 0 !!!
P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*No*) = 5/14

- Apply Laplacian correction (for all probabilities)

P(Outlook=*Overcast*|Play=*No*)=(0+1)/(5+3)=1/8

P(*No*|$\mathbf{x}'$): [P(*Overcast*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*)

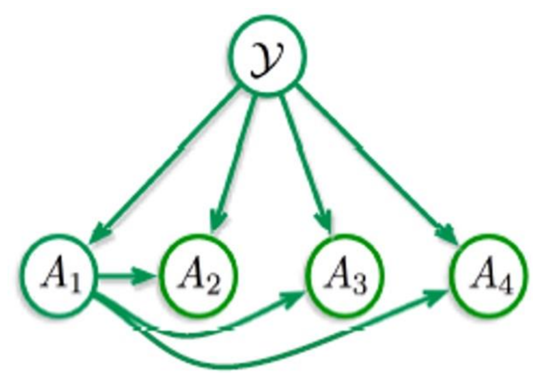
$$= \frac{0+1}{5+3} * \frac{1+1}{5+3} * \frac{4+1}{5+2} * \frac{3+1}{5+2} * \frac{5+1}{14+2} = 0.004783$$

P(*Yes*|$\mathbf{x}'$): $\frac{4+1}{9+3} * \frac{3+1}{9+3} * \frac{3+1}{9+2} * \frac{3+1}{9+2} * \frac{9+1}{14+2} = 0.01148 \Longrightarrow$ P(*Yes*|$\mathbf{x}'$) > P(*No*|$\mathbf{x}'$)

Independence Violation

MACQUARIE University

- For many real-world tasks, the independence assumption is violated
  - $P(X_1, \ldots, X_M|C) \neq P(X_1|C) \ldots P(X_M|C)$
- But, naïve Bayes classifier works surprisingly well
  - Why?
  - Model complexity is simpler (with fewer model parameters)
  - Avoid overfitting when training data sets are small
- Can relax the assumption by allowing certain types of dependence, e.g., AODE (Averaged One-Dependence Estimators)



NBC Pros and Cons

MACQUARIE University

- Pros
  - Very simple, and easy to implement.
  - Work well in practice even if NB assumption doesn't hold.
  - Highly scalable and fast, as it scales linearly with the number of features and data instances.
  - Can be used for both binary and multi-class classification.
  - Can make probabilistic predictions.
  - Can handle both continuous and discrete attributes.
  - Insensitive to irrelevant features.
- Cons
  - Strong assumption on NB conditional independence : any two features are independent given the output class.

Lecture Outline

- Probability Basics
- Naïve Bayes Classifier
- Practical

Naive Bayes Algorithms

MACQUARIE University

- Module sklearn.naive_bayes

GaussianNB	Gaussian Naive Bayes (GaussianNB).
CategoricalNB	Naive Bayes classifier for categorical features.
MultinomialNB	Naive Bayes classifier for multinomial models.
BernoulliNB	Naive Bayes classifier for multivariate Bernoulli models.
ComplementNB	The Complement Naive Bayes classifier described in Rennie et al. (2003).

GaussianNB

- *class* sklearn.naive_bayes.**GaussianNB**(**, priors=None, var_smoothing=1e-09*)
  - **priors:** probabilities of the classes. If specified, the priors are not adjusted according to the data.
- Attributes
  - **class_prior_:** probability of each class.
  - **theta_:** mean of each feature per class.
  - **var_:** variance of each feature per class.
- Demo
  - https://colab.research.google.com/drive/1UD-aholoiaLesQtsKesuJ4i7zxPS_KxN?usp=sharing

CategoricalNB

- *class* sklearn.naive_bayes.**CategoricalNB**(**, alpha=1.0, force_alpha=True, fit_prior=True, class_prior=None, ...*)
  - Suitable for classification with discrete features.
  - **alpha:** additive (Laplace(alpha=1)/Lidstone) smoothing parameter (no smoothing: set alpha=0 and force_alpha=True).
- Attributes
  - **class_log_prior_**: smoothed empirical log probability for each class. (why log?)
  - **feature_log_prob_**: Holds arrays of shape (n_classes, n_categories of respective feature) for each feature. Each array provides the empirical log probability of categories given the respective feature and class, $P(x_i|y)$. (why log?)

CategoricalNB

MACQUARIE University

- Demo
  - https://colab.research.google.com/drive/1UD-aholoiaLesQtsKesuJ4i7zxPS_KxN?usp=sharing

Summary

- Probability basic concepts
  - Joint/conditional/marginal probability
  - Expectation and variance
- Probabilistic modelling
- Optimal Bayes classifier
  - MAP decision rule
- Naïve Bayes classifier
  - Conditionally-independent assumption
  - Training: discrete and continuous probability estimation
  - Testing: calculate the probability
  - Zero probability issue