03_Naïve_Bayes

3.0 Why Naïve Bayes?

- Uncertainty
  - Can't conclude something with 100% confidence
- Weak Implications
  - Hard to establish concrete correlations between IF and THEN.
  - Handle vague associations.
- Imprecise Language
  - Natural language is ambiguous.
  - We describe facts with: sometimes, often, frequently, hardly, ...
  - Difficult to establish IF-THEN rules based on NL.

3.1 Basic Probability Theory

3.1.1 Probability 概率

- The probability of an event
  - = the proportion of cases in which the event occurs.
  - Expression: From 0 (absolute impossible) $\rightarrow$ Unity (Absolute certain).
  - Mostly strictly between 0 and 1. Each event has at least two outcomes: success or failure.
    - $P(\text{success}) = \dfrac{s}{s+f}$
    - $P(\text{failure}) = \dfrac{f}{s+f}$

3.1.2 Conditional Probability 条件概率

- Let: A, B be an Event.
- Conditional Probability of A over B:
  - The probability that: If B occur, then A occur.
  - $P(A|B) = \dfrac{\#.\,(A \text{ and } B \text{ occur})}{\#.\,(B \text{ occur})}$
  - $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B|A)P(A)}{P(B)}$ (yields the Bayesian Rule)

3.2 Bayesian Reasoning

3.2.1 Bayesian Rule

Given

- An event E.
  - E stands for "Evidence".
  Do
  Get the prob that event H (Hypothesis) will occur, as P.
- $P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H) \cdot P(H)}{\left[P(E|H) \cdot P(H)\right] + \left[P(E|\neg H) \cdot P(\neg H)\right]}$

3.2.2 Variances

Single Evidence, Multiple Hypothesis:

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{\sum_{k=1}^{m} \left[P(E|H_k) \cdot P(H_k)\right]}$$

Multiple Evidence, Multiple Hypothesis:

$$P(H_i|E_1, E_2, \ldots, E_n) = \frac{P(E_1, E_2, \ldots, E_n) \cdot P(H_i)}{\sum_{k=1}^{m} \left[P(E_1, E_2, \ldots, E_n|H_k) \cdot P(H_k)\right]}$$

- $\approx \frac{\left[P(E_1|H_i) \cdot P(E_2|H_i) \cdot \ldots \cdot P(E_n|H_i)\right] \cdot P(H_i)}{\sum_{k=1}^{m} \left[P(E_1|H_k) \cdot P(E_2|H_k) \cdot \ldots \cdot P(E_n|H_k) \times P(H_k)\right]}$, if conditional independence holds.
- $= \frac{P(H_i) \cdot \left[\prod_{a=1}^{n} P(E_a|H_i)\right]}{\sum_{k=1}^{m} \left[P(H_k) \cdot \prod_{b=1}^{n} P(E_b|H_k)\right]}$

Example

Given the prior an conditional probs as follows:

	H_1	H_2	H_3
$P(H_i)$	0.40	0.35	0.25
$P(E_1\|H_i)$	0.3	0.8	0.5

	H_1	H_2	H_3
$P(E_2\|H_i)$	0.9	0.0	0.7
$P(E_3\|H_i)$	0.6	0.7	0.9

- Want $P(H_1|E_3)$.
- $P(H_3|E_3) = \dfrac{P(E_3|H_3)P(H_3)}{P(E_3)}$, where
    - $P(E_3|H_3) \cdot P(H_3) = 0.9 \times 0.25 = 0.36$
    - $P(E_3) = \Big[P(E_3|H_1) \cdot P(H_1)\Big] \times \Big[P(E_3|H_2) \cdot P(H_2)\Big] \times \Big[P(E_3|H_3) \cdot P(H_3)\Big]$
        - $= 0.6 \times 0.4 + 0.7 \times 0.35 + 0.9 \times 0.25 = 0.2838$

3.3 Naïve Bayes Classifier

3.3.1 Maximum A Posteriori

$$H_{\text{conclusion}} = \text{argmax}_{h \in H} P(h|E)$$

- $= \text{argmax}_{h \in H} \dfrac{P(E|h) \cdot P(h)}{P(E)}$
- $= \text{argmax}_{h \in H} \Big[P(E|h) \cdot P(h)\Big]$

Omit the $P(E)$ since it's constant, which is independent from the hypothesis.

- Here, $P(E|h) \cdot P(h)$ is the Maximum A Posteriori.

3.3.2 Naïve Bayes Estimation

- Given:
    - A conjunctive test sample: $x_1, x_2, \ldots, x_n$
- $c_{MAP} = \text{argmax}_{c_j \in C} \Big[P(c_j|x_1, x_2, \ldots x_n)\Big]$
    - $= \text{argmax}_{c_j \in C} \left[\dfrac{P(x_1, x_2, \ldots, x_n|c_j) \cdot P(c_j)}{P(x_1, x_2, \ldots, x_n)}\right]$
    - $= \text{argmax}_{c_j \in C} \Big[P(x_1, x_2, \ldots, x_n|c_j) \cdot P(c_j)\Big]$
    - $= \text{argmax}_{c_j \in C} \Big[P(x_1|c_j) \times P(x_2|c_j) \times \ldots \times P(x_n|c_j) \times P(c_j)\Big]$
    - $= \text{argmax}_{c_j \in C} \Big[P(c_j) \times \prod_{k=1}^{n} P(x_k|c_j)\Big]$

Example

Day	Outlook	Temp	Humitity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Known: Outlook=sunny, Temp=cool, Humidity=high, Wind=strong.
Want: Play tennis or not?

Do:

- $MAP(\text{Yes}|\text{sunny}, \text{cool}, \text{high}, \text{strong})$
  - $= P(\text{sunny}, \text{cool}, \text{high}, \text{strong}|\text{Yes}) \times P(\text{Yes})$
  - $= P(\text{sunny}|\text{Yes}) \times P(\text{cool}|\text{Yes}) \times P(\text{high}|\text{Yes}) \times P(\text{strong}|\text{Yes}) \times P(\text{Yes})$
  - $= [\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9}] \times \frac{9}{14}$
  - $= 0.005291005291$
- $MAP(\text{NO}|\text{sunny}, \text{cool}, \text{high}, \text{strong})$
  - $= P(sunny, cool, high, strong|No) \times P(No)$
  - $= P(sunny|No) \times P(cool|No) \times P(high|No) \times P(strong|No) \times P(No)$
  - $= [\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5}] \times \frac{5}{14}$
  - $= 0.02057142857$

3.2.3 Add-1 Smoothing

Initially, we have:

$$c_{\text{target}} = \text{argmax}_{c_j \in C}\left[P(c_j)\prod_{i=1}^{n} P(x_i|c_j)\right]$$

We could observe that:

$$P(x_i|c_j) = \frac{\#.\,(x \in c_j \wedge x = x_i)}{\#.\,(x \in c_j)} = \frac{n_c}{n}$$

where the number of x that's in class c_j could be 0, yielding $P(x_i|c_j)$ to be 0.

What's worse, if $P(x_i|c_{j_1})$ becomes 0 for j_1, even if $P(x_i|c_{j_2})$ is very large for j_2, the entire $MAP = P(c_j)\prod_{i=1}^{n} P(x_i|c_j)$ would be still cast to 0.

Resolution: Add-1 smoothing, with:

- Prior: $P(c_j) = \dfrac{(\#.\,c \in C \wedge c = c_j) + m_{\text{prior}} \times p_{\text{prior}}}{(\#.\,c \in C) + m_{\text{prior}}}$, where $m_{\text{prior}} \in \mathbb{R}^+$ and $p_{\text{prior}} \in [0, 1]$
- Evidence: $P(x_i|c_j) = \dfrac{(\#.\,x_i \in c_j) + m_{\text{evid}} \times p_{\text{evid}}}{(\#.\,c_j) + m_{\text{evid}}}$, where $m_{\text{evid}} \in \mathbb{R}^+$ and $p_{\text{evid}} \in [0, 1]$

  The intuition is that:
- We create m imaginary prior examples, where a portion of p is the one to be examined.

3.2.4 Continuous x

Observations may be continuous. Use Gaussian Distribution instead.

$$P(x_i|c_k) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{\dfrac{-(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}} = \mathcal{N}(x_i, \mu_{ik}, \sigma_{ik})$$

That is, for a specific class c_j, extract all the values $x_i \in c_j$ and form a normal distribution. This determines two variables:

- μ, the mean/expectation
  - $\mu_{ik} = \dfrac{1}{n_k}\sum_{j=1}^{n_k} x_{ikj}$
- σ, the standard deviation

  - $\sigma_{ik} = \sqrt{\dfrac{1}{n_k}\sum_{j=1}^{n_k}(x_{ikj} - \mu_{ik})^2}$

  Note that:
- i - the i-th feature,
- k - the target class,
- j - within a class k, the j-th data.

  After the two variables are set, the probability $P(x_i|c_j)$ can be thus calculated.

- $P(x_i|c_j)$ means that:
  - Under the given class c_j, there's a lot of continuous numeric input under the i-th feature.
  - We are given a test value x_i
  - $P(x_i|c_j)$ gives us the probability that x_i is here according to all the numeric inputs mentioned above, with respect to a gaussian/normal distribution.

Example

There is a set of data, having two classes:

- $\text{Play} = \text{Yes}$
- $\text{Play} = \text{No}$
  Each data sample is composed of 5 features, listed with class-wise distribution below:

Day	Outlook	Temp	Humitity	Wind	Play Tennis
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	86	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	91	Strong	No

Classify for a test sample:

- $\text{Outlook} = \text{Sunny}, \text{Temp} = 66, \text{Humidity} = 90, \text{Wind} = \text{Strong}$

Calculate Prior probabilities.

- $P(\text{Yes}) = \dfrac{9}{14}$
- $P(\text{No}) = \dfrac{5}{14}$

Calculate Posterior probabilities.

Outlook

- Yes
  - $P(\text{Sunny}|\text{Yes}) = \dfrac{2}{9}$
- No
  - $P(\text{Sunny}|\text{Yes}) = \dfrac{3}{5}$

Temperature

- Yes
  - $\mu_{\text{temp,Yes}} = \dfrac{1}{9}\left[83 + 70 + 68 + 64 + 69 + 75 + 75 + 72 + 81\right] = 73$
  - $\sigma_{\text{temp,Yes}} = \sqrt{\dfrac{1}{9}\sum_{j=1}^{n_{\text{Yes}}}(x_{\text{temp,Yes},j} - 73)^2} = \dfrac{4\sqrt{19}}{3}$
  - $\implies P(66|\text{Yes}) = \dfrac{1}{\frac{4\sqrt{19}}{3}\cdot\sqrt{2\pi}} e^{\dfrac{-(66-73)^2}{2\cdot\frac{304}{9}}} = 0.0332$
- No
  - $\mu_{\text{temp,No}} = \dfrac{1}{5}\left[85 + 80 + 65 + 71 + 71\right] = 74.6$
  - $\sigma_{\text{temp,No}} = \sqrt{\dfrac{1}{5}\sum_{j=1}^{n_{No}}(x_{\text{temp,No,j}} - 74.6)^2} = \sqrt{49.84} = 7.0597$
  - $\implies P(66|\text{No}) = \dfrac{1}{7.0597 \times \sqrt{2\pi}} e^{\dfrac{-(66-74.6)^2}{2\times 49.84}} = 0.0269$

Humidity

- Yes
  - $\mu_{\text{humid,Yes}} = \dfrac{1}{9}\left[86 + 96 + 80 + 65 + 70 + 80 + 70 + 90 + 75\right] = \dfrac{712}{9}$
  - $\sigma_{\text{humid,Yes}} = \sqrt{\dfrac{1}{9}\sum_{j=1}^{n_{Yes}}(x_{\text{humid,Yes},j} - \frac{712}{9})^2} = \sqrt{92.7654} = 9.6315$
  - $\implies P(90|\text{Yes}) = \dfrac{1}{9.6315 \times \sqrt{2\pi}} e^{\dfrac{-(90-\frac{712}{9})^2}{2\times 92.7654}} = 0.0219$
- No

- $\mu_{\text{humid,No}} = \dfrac{1}{5}\left[95+90+70+95+91\right] = \dfrac{441}{5}$
- $\sigma_{\text{humid,No}} = \sqrt{\dfrac{1}{5}\sum_{j=1}^{n_{No}}(x_{\text{temp,No},j} - \frac{441}{9})^2} = \sqrt{75.76} = 8.7040$
- $\implies P(90|\text{No}) = \dfrac{1}{8.7040 \times \sqrt{2\pi}} e^{\dfrac{-\left(90 - \frac{441}{5}\right)^2}{2 \times 75.76}} = 0.0449$

Wind

- Yes
  - $P(\text{Strong}|\text{Yes}) = \dfrac{3}{9}$
- No
  - $P(\text{Strong}|\text{No}) = \dfrac{3}{5}$

Therefore, to sum up:

- $MAP(\text{Yes}|\text{Summy}, 66, 90, \text{Strong}) = \left(\dfrac{2}{9} \times 0.0332 \times 0.0219 \times \dfrac{3}{9}\right) \times \dfrac{9}{14} = 3.4623 \times 10^{-5}$
- $MAP(\text{No}|\text{Summy}, 66, 90, \text{Strong}) = \left(\dfrac{3}{5} \times 0.0269 \times 0.0449 \times \dfrac{3}{5}\right) \times \dfrac{5}{14} = 1.5529 \times 10^{-4}$

  Therefore, the classification result is No.