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 [DEF] Probability

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

3.1.2 [DEF] Conditional Probability

- Let: A, B: Event
- Conditional Probability:
  - The probability that: If B occur, then A occur.
  - $P(A|B) = \frac{num(AandBoccur)}{num(Boccur)}$

- $P(A|B)=\dfrac{P(A\cap B)}{P(B)}=\dfrac{P(B|A)P(A)}{P(B)}$ (yields the Bayesian Rule)

3.2 Bayesian Reasoning

3.2.1 Baysian Rule:

- Given: Event E (Evidence)
- Get: The prob. that event H (Hypothesis) will occur, as P.
  - $P(H|E)=\dfrac{P(E|H)P(H)}{P(E)}=\dfrac{P(E|H)P(H)}{P(E|H)P(H)+P(E|\neg H)P(\neg H)}$

3.2.2 Variances:

- Single Evidence, Multiple Hypothesis:
  - $P(H_i|E)=\dfrac{P(E|H_i)P(H_i)}{\sum_{k=1}^m P(E|H_k)P(H_k)}$
- Multiple Evidence, Multiple Hypothesis:
  - $P(H_i|E_1,E_2,\dots,E_n)=\dfrac{P(E_1,E_2,\dots,E_n)P(H_i)}{\sum_{k=1}^m P(E_1,E_2,\dots,E_n|H_k)P(H_k)}$
  - $\approx\dfrac{P(E_1|H_i)\times P(E_2|H_i)\times\dots\times P(E_n|H_i)\times P(H_i)}{\sum_{k=1}^m\left[P(E_1|H_k)\times P(E_2|H_k)\times\dots\times P(E_n|H_k)\times P(H_k)\right]}$, if conditional independence holds.
  - $=\dfrac{P(H_i)\prod_{a=1}^n P(E_a|H_i)}{\sum_{k=1}^m P(H_k)\prod_{b=1}^n P(E_b|H_k)}$

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
$P(E_2|H_i)$	0.9	0.0	0.7
$P(E_3|H_i)$	0.6	0.7	0.9
- Want $P(H_1	E_3)$.		
- $P(H_3	E_3)=\dfrac{P(E_3	H_3)P(H_3)}{P(E_3)}$	

	H_1	H_2	H_3
- $P(E_3	H_3)P(H_3)=0.9 \times 0.25=0.36$		
- $P(E_3)=P(E_3	H_1)P(H_1)\times P(E_3	H_2)P(H_2)\times P(E_3	H_3)P(H_3)$

```
- $=0.6\times 0.4+0.7\times 0.35+0.9\times 0.25=0.2838$
```

3.3 Naïve Bayes Classifier

3.3.1 Maximum A Posteriori

- $H_{conclusion} = argmax_{h \in H} P(h|E)$
  - $= argmax_{h \in H} \frac{P(E|h)P(h)}{P(E)}$
  - $= argmax_{h \in H} P(E|h)P(h)$
- Omit the $P(E)$ since it's constant, which is independent from the hypothesis.

3.3.2 Naïve Bayes Estimation

- Given:
  - A conjunctive test sample: $x_1, x_2, \ldots, x_n$
- $c_{MAP} = argmax_{c_j \in C} P(c_j|x_1, x_2, \ldots x_n)$
  - $= argmax_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n|c_j)P(c_j)}{P(x_1, x_2, \ldots, x_n)}$
  - $= argmax_{c_j \in C} P(x_1, x_2, \ldots, x_n|c_j)P(c_j)$
  - $= argmax_{c_j \in C} [P(x_1|c_j) \times P(x_2|c_j) \times \ldots \times P(x_n|c_j)] \times P(c_j)$
  - $= argmax_{c_j \in C} P(c_j) \times \prod_{k=1}^{n} P(x_k|c_j)$

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**

Day	Outlook	Temp	Humitity	Wind	Play Tennis
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:

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

3.2.3 Add-1 Smoothing

Initially, we have:

$$c_{target} = argmax_{c_j \in C}[P(c_j) \prod_{i=1}^{n} P(x_i|c_j)]$$

We could observe that:

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

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.

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

3.2.4 Continuous x

Observations may be continuous. Use Gaussian Distribution instead.

$$P(x_i|c_j) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}} = Gaussian(x_i, \mu_{ik}, \sigma_{ik})$$

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:

- σ, the standard deviation
  - $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$
- μ, the mean/expectation
  - $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
  After the two variables are set, the probability $P(x_i|c_j)$ can be thus calculated.