COMP [56]630– Machine Learning

Lecture 15 – Naïve Bayes

© Sathyanarayanan Aakur

Logistics

- Midterm on March 1, 2024
  - During class hours (available until Sunday for Distance section)
  - 50 minutes
  - 40 1.25-point questions
  - Topics:
    - Machine Learning basics
    - Linear Regression
    - Basis Functions
    - Logistic Regression
    - MLP
    - Deep Learning basics
    - CNN
    - LSTM/RNN

Different types of classifiers

- Discriminative
  - Model a classification rule directly
    - Eg. Perceptron, logistic regression
  - Model the probability of class memberships given input data
    - Eg. Neural Networks with cross entropy (log loss)
- Generative
  - Make a probabilistic model of data within each class
    - Eg. Naïve Bayes, model-based classifiers
- Probabilistic
  - Output is a probability measure on the likelihood of an example belonging to a specific class given a set of features/observations.

Probability Basics

- Prior, conditional and joint probability
  - Prior probability: $P(X)$

Example: the chances of rolling a "4" with a die

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

So the probability $= \frac{1}{6}$

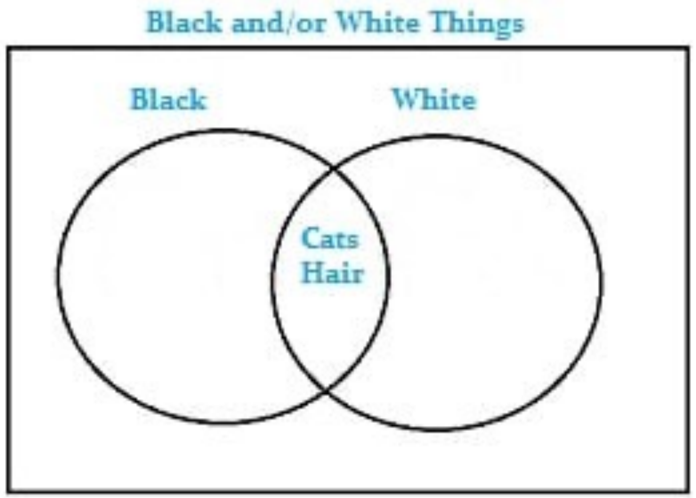
Probability Basics

- Prior, conditional and joint probability
  - Conditional probability: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
  - Conditional probability could describe an event like:
    - Event A is that it is raining outside, and it has a 0.3 (30%) chance of raining today.
    - Event B is that you will need to go outside, and that has a probability of 0.5 (50%).
  - A conditional probability would look at these two events in relationship with one another, such as the probability that it is both raining *and* you will need to go outside.

© Sathyanarayanan Aakur

Probability Basics

- Prior, conditional and joint probability
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Joint probability factors the likelihood of both events occurring.
    - Joint probability can also be described as the probability of the intersection of two (or more) events. The intersection can be represented by a Venn diagram:



© Sathyanarayanan Aakur

Probability Basics

- Prior, conditional and joint probability
  - Prior probability: $P(X)$
  - Conditional probability: $P(X_1 | X_2), P(X_2 | X_1)$
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship: $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(X_1 | X_2)P(X_2)$
  - Independence: $P(X_2 | X_1) = P(X_2), P(X_1 | X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- Bayesian Rule

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

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

© Sathyanarayanan Aakur

Probabilistic Classification

- Establishing a probabilistic model for classification
  - Discriminative model

$$P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$

  - Generative model

$$P(\mathbf{X} \mid C) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$

- MAP classification rule
  - MAP: Maximum A Posterior
  - Assign x to c^* if $P(C = c^* \mid \mathbf{X} = \mathbf{x}) > P(C = c \mid \mathbf{X} = \mathbf{x}) \quad c \neq c^*, \; c = c_1, \cdots, c_L$
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert: $P(C \mid \mathbf{X}) = \dfrac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})} \propto P(\mathbf{X} \mid C)P(C)$

Naïve Bayes

- Bayes classification

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_n \mid C)P(C)$$

Difficulty: learning the joint probability $P(X_1, \cdots, X_n \mid C)$

- Naïve Bayes classification
  - Making the assumption that all input attributes are independent

$$\begin{aligned} P(X_1, X_2, \cdots, X_n \mid C) &= \underline{P(X_1 \mid X_2, \cdots, X_n; C)} P(X_2, \cdots, X_n \mid C) \\ &= \underline{P(X_1 \mid C)}\, \underline{P(X_2, \cdots, X_n \mid C)} \\ &= P(X_1 \mid C) \underline{P(X_2 \mid C) \cdots P(X_n \mid C)} \end{aligned}$$

  - MAP classification rule

$$[P(x_1 \mid c^*) \cdots P(x_n \mid c^*)]P(c^*) > [P(x_1 \mid c) \cdots P(x_n \mid c)]P(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Naïve Bayes

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set **S**,

$$\text{For each target value of } c_i \; (c_i = c_1, \cdots, c_L)$$
$$\hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } \mathbf{S};$$
$$\text{For every attribute value } a_{jk} \text{ of each attribute } x_j \; (j = 1, \cdots, n; \; k = 1, \cdots, N_j)$$
$$\hat{P}(X_j = a_{jk} | C = c_i) \leftarrow \text{estimate } P(X_j = a_{jk} | C = c_i) \text{ with examples in } \mathbf{S};$$

Output: conditional probability tables; for $x_j,\ N_j \times L$ elements

  - Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_n)$,

Look up tables to assign the label c^* to $\mathbf{X}'$ if

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_n | c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_n | c)]\hat{P}(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Example 1

- Example: Play Tennis

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

© Sathyanarayanan Aakur

Example 1

- Learning Phase

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

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

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

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

$P(\text{Play}=Yes) = 9/14$ $\quad$ $P(\text{Play}=No) = 5/14$

Example 1

- Test Phase
  - Given a new instance,

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up 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

  - MAP rule

P(*Yes*|$\mathbf{x}'$): [P(*Sunny*|Yes)P(*Cool*|*Yes*)P(*High*|Yes)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

P(*No*|$\mathbf{x}'$): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|$\mathbf{x}'$) < P(*No*|$\mathbf{x}'$), we label $\mathbf{x}'$ to be "*No*".

Example 2

- Text classification with Naïve Bayes.
- Let us say we have the following data:

Document	Text	Class
1	I loved the movie	+
2	I hated the movie	-
3	a great movie. good movie	+
4	poor acting	-
5	great acting. a good movie	+

Example 2

- Step 1: Create variables from vocabulary.
  - In this data, we have 10 unique words:

  `< I, loved, the, movie, hated, a, great, poor, acting, good >`

Example 2

- Step 2: Create feature set.
  - Create frequency-based features for each variable/word

	a	acting	good	great	hated	i	loved	movie	poor	the
Doc0	0	0	0	0	0	1	1	1	0	1
Doc1	0	0	0	0	1	1	0	1	0	1
Doc2	1	0	1	1	0	0	0	2	0	0
Doc3	0	1	0	0	0	0	0	0	1	0
Doc4	1	1	1	1	0	0	0	1	0	0

© Sathyanarayanan Aakur

Example 2

- Step 3: Transform feature set based on class composition

Positive Class

	a	acting	good	great	hated	i	loved	movie	poor	the	class
Doc0	0	0	0	0	0	1	1	1	0	1	+
Doc2	1	0	1	1	0	0	0	2	0	0	+
Doc4	1	1	1	1	0	0	0	1	0	0	+

Negative Class

	a	acting	good	great	hated	i	loved	movie	poor	the	class
Doc1	0	0	0	0	1	1	0	1	0	1	-
Doc3	0	1	0	0	0	0	0	0	1	0	-

© Sathyanarayanan Aakur

Example 2

- Step 4: Computing probabilities.

$$P(+) = \frac{3}{5} = 0.6$$

- Step 5: Compute conditional probabilities:

$$P(w_k \mid +) = \frac{n_k + 1}{n + \mid \text{vocabulary}\mid}$$

Example 2

- Conditional Probabilities for + class:

$$P(I \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(loved \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(the \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(movie \mid +) = \frac{4+1}{14+10} = 0.20833$$

$$P(a \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(a \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(great \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(acting \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(good \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(poor \mid +) = \frac{0+1}{14+10} = 0.0417$$

$$P(hated \mid +) = \frac{0+1}{14+10} = 0.0417$$

Example 2

- Conditional Probabilities for - class:

$$P(-) = \frac{2}{5} = 0.4$$

$$P(I \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(loved \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(the \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(movie \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(a \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(great \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(acting \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(good \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(poor \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(hated \mid -) = \frac{1+1}{6+10} = 0.125$$

© Sathyanarayanan Aakur

Example 2

- Step 6: Inference for a given example
  - Input: “I hated the poor acting”
  - Hence x_i={I, hated, the, poor, acting}
  - Compute $P(+|x_i)$ and $P(-|x_i)$
    - $P(+|x_i) = P(+)*P(I \mid +)*P(hated \mid +)*P(the \mid +)*P(poor \mid +)*P(acting \mid +) = 6.03 \times 10^{-7}$
    - $P(-|x_i) = P(-)*P(I \mid -)*P(hated \mid -)*P(the \mid -)*P(poor \mid -)*P(acting \mid -) = 1.22 \times 10^{-5}$
  - $P(+|x_i) < P(-|x_i)$
    - => given example is negative!

Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks, $P(X_1,\cdots,X_n|C) \neq P(X_1|C)\cdots P(X_n|C)$
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the attribute value $X_j = a_{jk}$, $\hat{P}(X_j = a_{jk}|C = c_i) = 0$
  - In this circumstance, $\hat{P}(x_1|c_i)\cdots\hat{P}(a_{jk}|c_i)\cdots\hat{P}(x_n|c_i) = 0$ during test
  - For a remedy, conditional probabilities estimated with

$$\hat{P}(X_j = a_{jk}|C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)

Relevant Issues

- Continuous-valued Input Attributes
  - Numberless values for an attribute
  - Conditional probability modeled with the normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of attribute values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of attribute values X_j of examples for which $C = c_i$

  - Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_n),\ C = c_1, \cdots, c_L$

    Output: $n \times L$ normal distributions and $P(C = c_i)\ \ i = 1, \cdots, L$
  - Test Phase: for $\mathbf{X}' = (X'_1, \cdots, X'_n)$
    - Calculate conditional probabilities with all the normal distributions
    - Apply the MAP rule to make a decision

© Sathyanarayanan Aakur

Example for NB with continuous data

No.	y	x1	x2
1	KFC	180	75
2	KFC	165	61
3	McD	167	62
4	KFC	178	63
5	KFC	174	69
6	KFC	166	60
7	McD	167	59
8	McD	165	60
9	KFC	173	68
10	KFC	178	71
	?	177	72

Data from 10 engineers: their height (cm) and weight (kg), and their favorite fast food (KFC or McD)

Training

Calculate Prior Probabilities

$$p(y) = \begin{cases} 7/10 & \text{if } y = \text{KFC} \\ 3/10 & \text{if } y = \text{McD} \end{cases}$$

$$p(x_1|y = \text{KFC}) \qquad p(x_2|y = \text{KFC})$$

Calculate Conditional Probabilities

$$p(x_1|y = \text{McD}) \qquad p(x_2|y = \text{McD})$$

Training

Estimate the mean:

$$\mu = \frac{1}{6}(180 + 165 + 178 + 174 + 166 + 173 + 178) = 173$$

Estimate the squared standard deviation:

$$\sigma^2 = \frac{1}{5}[(180 - 173)^2 + (165 - 173)^2 + (178 - 173)^2 + (174 - 173)^2 + (166 - 173)^2 + (173 - 173)^2 + (178 - 173)^2] = 35$$

$$p(x_1 | y = \text{KFC}) = \frac{1}{\sqrt{2\pi(35)}} \exp(-\frac{(x_1 - 173)^2}{2(35)})$$

Training

$$\mu = \frac{1}{6}(75 + 61 + 63 + 69 + 60 + 68 + 71) = 67$$

$$\sigma^2 = \frac{1}{5}[(75 - 67)^2 + (61 - 67)^2 + (63 - 67)^2 + (69 - 67)^2 + (60 - 67)^2 + (68 - 67)^2 + (71 - 67)^2] = 31$$

$$p(x_2|y = \text{KFC}) = \frac{1}{\sqrt{2\pi(31)}} \exp(-\frac{(x_1 - 67)^2}{2(31)})$$

Training

$$\mu = \frac{1}{3}(167 + 167 + 165) = 166$$

$$\sigma^2 = \frac{1}{2}[(167 - 166)^2 + (167 - 166)^2 + (165 - 166)^2]$$
$$= 1.33$$

$$p(x_1|y = \mathrm{McD}) = \frac{1}{\sqrt{2\pi(1.33)}} \exp(-\frac{(x_1 - 166)^2}{2(1.33)})$$

Training

$$\mu = \frac{1}{3}(62 + 59 + 60) = 60$$

$$\sigma^2 = \frac{1}{2}[(62 - 60)^2 + (59 - 60)^2 + (60 - 60)^2]$$
$$= 2.33$$

$$p(x_2 | y = \mathrm{McD}) = \frac{1}{\sqrt{2\pi(2.33)}} \exp(-\frac{(x_1 - 60)^2}{2(2.33)})$$

Inference

- We want to find the probability of a co-worker's favourite food being a KFC, knowing that he is 177cm tall and weighs 72kg.

$$p(y = \text{KFC} | x_1 = 177, x_2 = 72)$$

- How?

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

Inference

$$p(y = \text{KFC}|x_1 = 177, x_2 = 72) \quad = \frac{p(x_1 = 177, x_2 = 72|y = \text{KFC}) \cdot p(y = \text{KFC})}{p(x_1 = 177, x_2 = 72)}$$

$$= \frac{p(x_1 = 177, x_2 = 72|y = \text{KFC}) \cdot p(y = \text{KFC})}{\sum_{i=\text{KFC,McD}} p(x_1 = 177, x_2 = 72|y = i) \cdot p(y = i)}$$

$$= \frac{p(x_1 = 177, x_2 = 72|y = \text{KFC}) \cdot p(y = \text{KFC})}{\sum_{i=\text{KFC,McD}} p(x_1 = 177, x_2 = 72|y = i) \cdot p(y = i)}$$

$$= \frac{p(x_1 = 177|y = \text{KFC}) \cdot p(x_2 = 72|y = \text{KFC}) \cdot p(y = \text{KFC})}{\sum_{i=\text{KFC,McD}} p(x_1 = 177|y = i) \cdot p(x_2 = 72|y = i) \cdot p(y = i)}$$

Inference

$$p(y = \text{KFC}|x_1 = 177, x_2 = 72) \quad = \frac{(0.0532)(0.0400)(\frac{7}{10})}{(0.0532)(0.0400)(\frac{7}{10}) + (0)(0)(\frac{3}{10})} \quad = 1$$

Conclusions

- Naïve Bayes based on the independence assumption
  - Training is very easy and fast; just requiring considering each attribute in each class separately
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more…