Bayesian inference and naïve Bayes

Akito Sakurai

Contents

- Bayes Theorem
- MAP and ML
- Bayes optimal classifier and Gibbs Algorithm
- Naïve Bayes

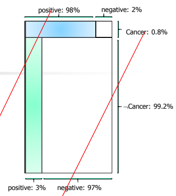
Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)\,P(A)$$

The definition of the conditional probability

An example
(Mitchell Chap. 6.2)



When a patient was screaned with a cancer test, the result was positive.
Does this patient have really cancer ?
The test reports positive 98% when in fact the patient has cancer; negative 97% negative when the patient does not have cancer.

Only 0.8% of whole population have cancer.

P(cancer) =	.008	P(¬cancer) =	.992
P(+ | cancer) =	.98	P(– | cancer) =	.02
P(+ | ¬cancer) =	.03	P(– | ¬cancer) =	.97

P(+) = P(+ | c'r) P(c'r) + P(+ | ¬c'r) P(¬c'r) = .0376

$$P(\text{cancer} \mid +) = \frac{P(+ \mid cancer)\,P(cancer)}{P(+)} = \mathbf{.209}$$

An example (Mitchell Exercise 6.1)

When a second test (statistically independent from the first one) was conducted, the result was positive. What is the posterior probability to be cancer?

P(cancer) =	.008	P(¬cancer) =	.992
P(+ | cancer) =	.98	P(– | cancer) =	.02
P(+ | ¬cancer) =	.03	P(– | ¬cancer) =	.97

$P(+_1+_2) = P(+_1+_2 \mid c'r)\,P(c'r) + P(+_1+_2 \mid \neg c'r)\,P(\neg c'r) = .00858$

$$P(\text{cancer} \mid +_1+_2) = \frac{P(+_1 +_2 \mid cancer)\,P(cancer)}{P(+_1 +_2)} = \mathbf{.896}$$

Basic Probability Formulas

Product rule (definition of conditional prob.):

$$P(A \wedge B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

Sum rule:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Theorem of total probability (if event Ai is mutually exclusive):

$$P(B) = \sum_{i=1}^{n} P(B, A_i) = \sum_{i=1}^{n} P(B \mid A_i)\,P(A_i)$$

What Bayes theorem tells us

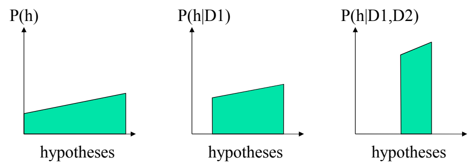
$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

P(h) = prior probability of a hypothesis h
P(D) = probability of a training dataset D
P(h|D) = posterior probability of h when D is given
P(D|h) = probability of D when h is given

We can chose more plausible hypothesis h that could produce the dataset D

Note: conditional probability does not reflect any causal relations if any.
Note: Can we think of "probability of hypotheses" ?

Development of posterior probability (when without noise)



Bayes Classifiers

Assumption: A training set consists of instances of different classes c_j described as a conjunction of attribute values

Task: Classify a new instance *d* based on a conjunction of attribute values into one of the classes $c_j \in C$

Key idea: Assign the most probable class c_{MAP} using Bayes Theorem.

$$\begin{aligned} c_{MAP} &= \operatorname*{argmax}_{c_j \in C} P(c_j \mid x_1, x_2, \ldots, x_n) \\ &= \operatorname*{argmax}_{c_j \in C} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)} \\ &= \operatorname*{argmax}_{c_j \in C} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j) \end{aligned}$$

Desirable Properties of Bayes Classifier

- *Combines prior knowledge and observed data:* prior probability of a hypothesis multiplied with probability of the hypothesis given the training data
- *Incremental:* with each training example, the prior and the likelihood can be updated dynamically: flexible and robust to errors.
- *Probabilistic hypothesis:* outputs not only a classification, but a probability distribution over all classes

Contents

- Bayes Theorem
- MAP and ML
- Bayes optimal classifier and Gibbs Algorithm
- Naïve Bayes

Maximum A Posterior

- Based on Bayes Theorem, we can compute the *Maximum A Posterior* (MAP) hypothesis for the data
- We are interested in the best hypothesis in some space H given observed training data D.

$$\begin{aligned} h_{MAP} &\equiv \operatorname*{argmax}_{h \in H} P(h \mid D) \\ &= \operatorname*{argmax}_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} \\ &= \operatorname*{argmax}_{h \in H} P(D \mid h) P(h) \end{aligned}$$

H: set of all hypothesis.

Note: We can drop *P*(*D*) as the probability of the data is constant (and independent of the hypothesis).

Maximum Likelihood

- Now assume that all hypotheses are equally probable a priori, i.e., $P(h_i) = P(h_j)$ for all h_i, h_j belonging to H.
- This corresponds to assuming a *uniform prior*. It simplifies computing the posterior (is it OK?):

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

- This hypothesis is called the *maximum likelihood hypothesis.*

An interpretation of ML

- In the real world, prior distribution is believed to be unknown, incomputable, or non-existent
  - For example, is there a prior distribution of words in some document? If it exists, it can vary depending on age, social background, population distribution, etc.
- If the prior does not exist, maximizing likelihood is a natural way to think of.

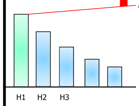
ML is equivalent to MAP when the hypotheses distribute uniformly, i.e., equivalent to suppose that the prior is uniform.
Is this valid?

Contents

- Bayes Theorem
- MAP and ML
- Bayes optimal classifier and Gibbs Algorithm
- Naïve Bayes

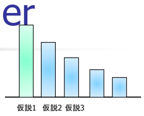
The most probable class

- We have sought the most likely under samples D hypothesis (e.g.: h_{MAP}).
- What about the most likely (probable) classification (class)?



- $h_{MAP}(x)$ is not most probable !
  - In the following, what is the most probable class of x ?
  - 3 hypotheses: $P(h_1|D)$=0.4, $P(h_2|D)$=0.3, $P(h_3|D)$=0.3
  - A new sample: $h_1(x)$=+, $h_2(x)$=–, $h_3(x)$= –

Bayes optimal classifier



$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D)$$

Note: Bayes optimal classifier is not necessarily in H.

Note: Some papers report that this works well, but in some cases we tried, the result was no better than MAP and ML. It is very interesting to know when it works and when it does not.

Note: Is it feasible? It looks to take long time to calculate.

An example (Mitchell Chap. 6.7)

$P(h_1 \mid D) = .4$	$P(-\mid h_1) = 0$	$P(+\mid h_1) = 1$
$P(h_2 \mid D) = .3$	$P(-\mid h_2) = 1$	$P(+\mid h_2) = 0$
$P(h_3 \mid D) = .3$	$P(-\mid h_3) = 1$	$P(+\mid h_3) = 0$

Therefore:

$$\sum_{h_i \in H} P(+ \mid h_i) P(h_i \mid D) = .4$$

$$\sum_{h_i \in H} P(- \mid h_i) P(h_i \mid D) = .6$$

And:

$$\arg\max_{c_j \in \{+,-\}} \sum_{h_i \in H} P(c_j \mid h_i) P(h_i \mid D) = -$$

Gibbs classifier

1. Select a hypothesis randomly according to P(h|D)
2. Classify a new example following the h

Good news: If a hypothesis is randomly sampled from P(h) ,

$E[error_{Gibbs}] \le 2E[error_{BayesOptimal}]$

(See "Mitchell Machine Learning Chap. 6.8")

Effective when there are so many hypothesis that a Bayes optimal is hard to calculate and we repeat the inferences

Contents

- Bayes Theorem
- MAP and ML
- Bayes optimal classifier and Gibbs Algorithm
- Naïve Bayes

Naïve Bayes classifier

- Since (although?) simple, it is wellknown
  - More accurate than expected, although simple
  - Fast as is expected, since simple
- Bayes Theorem + Assumption *conditional independence*
  - The assumption hardly holds in the real world
  - In the real world, though, it works well
- Successful applications:
  - Text classification,
  - Diagnosis, and many others

Naïve Bayes is not a Bayesian

Difficulty in Bayes classifier

- Recall that for a set of attributes $<a_1,\ldots,a_n>$ of x, to infer the class that x belongs

$$c_{MAP} = \arg\max_{c_j \in C} P(c_j \mid a_1, a_2, \ldots, a_n)$$
$$= \arg\max_{c_j \in C} \frac{P(a_1, a_2, \ldots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \ldots, a_n)}$$
$$= \arg\max_{c_j \in C} P(a_1, a_2, \ldots, a_n \mid c_j) P(c_j)$$

- Difficulty: Huge dataset is required to infer $P(a_1\ldots a_n|c_j)$, since there are huge number of parameters $(\Pi|A_i|)$ (for two value attributes, 2^n parameters for n attributes)

Naïve Bayes classifier

- Naïve Bayes assumption: attributes are mutually independent when the class is given
  - $P(a_1,\ldots,a_n|c_j) = P(a_1|c_j)\, P(a_2|c_j)\, \ldots\, P(a_n|c_j)$
  - *conditional independence* (given the class)
  - Reduces the number of parameters to infer:

    $\Pi|A_i|$ (=O(2^n)) $\rightarrow \Sigma|A_i|$ (=O(n))
- Under this assumption, c_{MAP} becomes

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(a_i \mid c_j)$$

Naïve Bayes: an Algorithm

Training (for a set of instances)

Estimate the probability that an instance x belongs to a class c_j

$P^{\wedge}(c_j) = P(c_j)$'s estimator

Estimate the probability that the i –th attribute value of an instance x belonging to the class c_j is a_i .

$P^{\wedge}(a_i|c_j) = P(a_i|c_j)$'s estimator

Class(x)

$$c_{NB} = \arg\max_{c_j \in C} \hat{P}(c_j) \prod_i \hat{P}(a_i \mid c_j)$$

Naïve Bayes: Estimation

- How can we estimate $P(c_j)$ and $P(a_i|c_j)$?
  - A standard method that statistics tells us
    - Use frequency of the samples
    - $P(c)$ is estimated by $\mathrm{count}(c) / N$
    - $P(A|B)$ is estimated by $\mathrm{count}(A \wedge B) / \mathrm{count}(B)$
  - Example: 100 samples. 70 + and 30 –
    - P(+)=0.7 and P(-)=0.3
    - Among 70 positives, in 35 a_1=SUNNY
    - P(a_1=SUNNY|+)=0.5

Example: Play Tennis

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Overcast	Hot	High	No	Yes
Rainy	Mild	High	No	Yes
Rainy	Cool	Normal	No	Yes
Rainy	Cool	Normal	Yes	No
Overcast	Cool	Normal	Yes	Yes
Sunny	Mild	High	No	No
Sunny	Cool	Normal	No	Yes
Rainy	Mild	Normal	No	Yes
Sunny	Mild	Normal	Yes	Yes
Overcast	Mild	High	Yes	Yes
Overcast	Hot	Normal	No	Yes
Rainy	Mild	High	Yes	No

There are two classes: to play tennis (Play=Yes) and not to play tennis (Play=No)

Please infer if on the following day they played tennis or not

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

From Tom Mitchell's Machine Learning

A solution

- For the *PlayTennis* , and a new instance
  <Outlook=sunny, Temp=cool, Humid=high, Windy=true>
- We want to calculate:

$$c_{NB} = \arg\max_{c_j \in C} \hat{P}(c_j) \prod_i \hat{P}(a_i | c_j)$$

- $\hat{P}(Y)\hat{P}(sunny|Y)\hat{P}(cool|Y)\hat{P}(high|Y)\hat{P}(true|Y) = 0.0053$
  $\hat{P}(N)\hat{P}(sunny|N)\hat{P}(cool|N)\hat{P}(high|N)\hat{P}(true|N) = 0.0206$

$$\Rightarrow c_{NB} = No$$

Procedure

Divided samples into the classes

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

Procedure

Count and infer parameters

	A1=Outlook		A2=Temperature		A3=Humidity		A4=Windy	
counts	Sunny	2	Hot	2	High	3	False	6
	Overcast	4	Mild	4	Normal	6	True	3
	Rainy	3	Cool	3				
	sum	9	sum	9	sum	9	sum	9
probs	Sunny	2/9	Hot	2/9	High	3/9	False	6/9
	Overcast	4/9	Mild	4/9	Normal	6/9	True	3/9
	Rainy	3/9	Cool	3/9				

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

	A1=Outlook		A2=Temperature		A3=Humidity		A4=Windy	
counts	Sunny	3	Hot	2	High	4	False	2
	Overcast	0	Mild	2	Normal	1	True	3
	Rainy	2	Cool	1				
	sum	5	sum	5	sum	5	sum	5
probs	Sunny	3/5	Hot	2/5	High	4/5	False	2/5
	Overcast	0/5	Mild	2/5	Normal	1/5	True	3/5
	Rainy	2/5	Cool	1/5				

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

Procedure

Combine them to make one table

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

p(m) should not be forgotten

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Rainy	Cool	Normal	True	No
Sunny	Mild	High	False	No
Rainy	Mild	High	True	No

A1=Outlook			A2=Temperature			A3=Humidity			A4=Windy			m=Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Procedure

Inference

$$p(m_j \mid x) = p(x \mid m_j)\, p(m_j) / p(x)$$
$$= p(a_1, \ldots, a_n \mid m_j)\, p(m_j) / p(x)$$
$$= \left(\prod_{i=1}^{n} p(a_i \mid m_j) \right) p(m_j) \Big/ p(x)$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

A new unseen x

p(Play=yes | x)
= p(Outlook=Sunny | Play=yes)
 * p(Temp=Cool | Play=yes)
 * p(Humidity=High | Play=yes)
 * p(Windy=True | Play=yes)
 * p(Play=yes) / p(x)
= (2/9) * (3/9) * (3/9) * (3/9)
 *(9/14) / p(x)
= 0.0053 / p(x)

p(Play=no | x)
= p(Outlook=Sunny | Play=no)
 * p(Temp=Cool | Play=no)
 * p(Humidity=High | Play=no)
 * p(Windy=True | Play=no)
 * p(Play=no) / p(x)
= (3/5) * (1/5) * (4/5) * (3/5)
 *(5/14) / p(x)
= 0.0206 / p(x)

In other words p(Play=yes | x) < p(Play=no | x)
i.e., "they did not (will not) play tennis"

Note: It is clear that you do need to consider 1/p(x) , which is common among all alternatives.

Naïve Bayes: Conditional independence

- Is it necessary?
- What happens if the assumption does not hold?
  - i.e. if $P(a_1,\ldots,a_n|c_j) \neq P(a_1|c_j)\, P(a_2|c_j)\ldots P(a_n|c_j)$
- If the following (weak) condition holds, the prediction is the same as Bayes classifier:

$$\arg\max_{c_j \in C} P(a_1 \mid c_j) P(a_2 \mid c_j) \ldots P(a_n \mid c_j) P(c_j)$$
$$= \arg\max_{c_j \in C} P(a_1, a_2, \ldots, a_n \mid c_j) P(c_j)$$

- But, the probability obtained in the prediction happens to be unrealistically close to 0 or 1

Naïve Bayes: a Problem

- What happens if an attribute value a_i is not observed for a class c_j ?
  - Estimator of $P(a_i|c_j)=0$ because count$(a_i \wedge c_j) = 0$
  - Big impacts: if this is 0, any products are 0 !
- A solution: use Laplace correction.
  - $\hat{P}(a_i \mid c_j) = \dfrac{n_c + mp}{n + m}$
  - n : # of training samples for $c = c_j$
  - n_c : # of training samples for $c = c_j$ and $a = a_j$
  - p : prior probability (estimator) $P\hat{}(a_i|c_j)$ (uniform distribution is common)
  - m : pseudo-count (commonly the number of attribute values)

 m=1 is another choice which works better in many cases, too

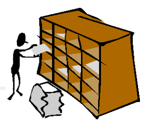
Note: Laplace correction

- (in parameter estimations from frequency) supposing a prior distribution for the parameter, obtain a MAP estimator.
- Beta distribution is the prior:
 $f(x;\alpha,\beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$
- The posterior mean of the parameter is the Laplace correction. If the likelihood is a result of a Bernoulli trial:
 $\hat{\theta} = (n_0 + \alpha)/\big((n_0 + n_1) + \alpha + \beta\big)$

Note: smoothing

- In an estimation of statistical model, assigning a small probability to events that did not occur is called smoothing
 http://www.jaist.ac.jp/project/NLP_Portal/doc/glossary/index.html
- In natural language processing, frequencies of a word or a sequence of n words (n-gram) are often used. When n grows, n-gram becomes scarce, i.e., many n-grams do not occur. To solve the problem many techniques were invented.
  - Laplace smoothing (additive smoothing)
  - Linear interpolation
  - Good-Turing smoothing
  - Katz smoothing
  - Church-Gale smoothing
  - Witten-Bell smoothing
  - Kneser-Ney smoothing
  - ....
  - Hierarchical Pitman-Yor language model

Text classification



- Texts classification:
  - Classifying documents (mail, news, web pages, etc. or a paragraph, a sentence, etc.)
  - Classifying e-mails into spam or not.
  - Classifying blogs into splog or not
  - Classifying news into interesting or not (to a person)
  - Classifying reviews of a product into groups of good reputation or not
  - Classifying reviews into trustable or not
  - Classifying open ended questions for questionnaire surveys
  - Classifying Q and A's at a call-center.
- Naïve Bayes works well
  - How to apply Naïve Bayes ?
  - Point: How to represent a sample (i.e. document), attributes?

Document representation

Bag-of-words

- Document as a vector of frequency of words in it
  - "Bag" implies discarding positions where the word occurs, and
  - disregarding the sequences (contexts) of a word
  - i.e. if keio, gijuku, and university are words, there would be no difference between keio gijuku university, keio university giju, and gijuku keio university
- "what are words" is important, which should not differ among documents.
  - In English, "dog" and "dogs" should be treated as the same
- Ignore words not relevant to classification
  - In Japanese, particles such as ha, ga, mo, ya, etc are the ones
  - In English, prepositions
  - The words that have syntactic function but have no meaning are called functional words.
- Ignore words that are close to noise
  - Very low frequent words such as appearing just once.

Document representation (cntd.)

- Representation itself is like naïve Bayes
  - Because representation is not inference, it is not naïve Bayes, but it really looks like naïve Bayes.
  - Probability of the occurrence of a document is formulated in naïve Bayes fashion.
  - Suppose that for each class of documents, the probability that a specific word occurs in a document is known as $P(w_1|c_j)$, $P(w_2|c_j)$, …, $P(w_n|c_j)$. If $w_1, w_2, \ldots, w_n$ are the words that occur in a document, then the probability that the document occurs is

$$P(\text{doc}|c_j)=P(w_1|c_j)^{\text{TF}(w1)}\, P(w_2|c_j)^{\text{TF}(w2)} \ldots P(w_n|c_j)^{\text{TF}(wn)}$$

where TF(w) is the term frequency of a word w in a document doc

出現確率をこう書けば naïve Bayes といえよう

Document classification by Naïve Bayes

- For a document doc,

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{w_k \in Voc} P(w_k \mid c_j)^{TF(w_k, doc)}$$

where $TF(w_k,\text{doc})$ = frequency of w_k in doc and Voc is a set of all the words that we consider

- To represent word frequencies in a document, we need Laplace correction. The following estimator is used; where n_j=the number of words in a class c_j , $n_{k,j}$=the number of occurrences of word w_k in class c_j .

$$P(w_k \mid c_j) = \frac{n_{k,j}+1}{n_j + |Voc|}$$

Twenty News Groups (Joachims 1996)

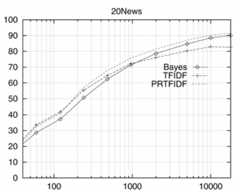
- 1000 training documents in each group
- Assign new documents to one of newsgroups

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x & rec.sport.hockey	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

T. Joachims. *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.* In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997, pp.143--151.

Twenty News Groups (Joachims 1996)

- Naive Bayes: 89% accuracy of classification
  - Highly frequent 100 word (the and of …) are deleted
    - The words such as functional words, words relatively useless for classification are categorized as stop words and are deleted
  - The words occurring less than 3 times are deleted
  - The words remained: 38,500



Note: the accuracy is overly high.. In every text in 20 Newsgroups has a "subject" field which is very helpful for classification. Although the subject field is now deleted, in the previous works the field might be utilized.

Accuracy vs. # of training data (1/3 is reserved for test)

Summary: Bayes inference and NB

$$P(h \mid D) = \frac{P(D \mid h)\,P(h)}{P(D)}$$

- Bayes inference:
  - ML: maximization of P(D|h)
  - MAP: maximization of P(h|D) ∝ P(D|h) P(h)
  - Posterior mean:
  - Bayes optimal: P(c|D) = ∫ P(c|h)P(h|D) dh
    - Supposing that the hypothesis distributes
- Naive Bayes: unrealistic assumption but works well in real world
  - Test classification is a good example

Exercise

- Apply Naïve Bayes to the following training data (left) and the test data (right) to find out the class label (skiing). Note: Apply Laplace correction.

snow	weather	season	physical condition	go skiing
sticky	foggy	low	rested	no
fresh	sunny	low	injured	no
fresh	sunny	low	rested	yes
fresh	sunny	high	rested	yes
fresh	sunny	mid	rested	yes
frosted	windy	high	tired	no
sticky	sunny	low	rested	yes
frosted	foggy	mid	rested	no
fresh	windy	low	rested	yes
fresh	windy	low	rested	yes
fresh	foggy	low	rested	yes
fresh	foggy	low	rested	yes
sticky	sunny	mid	rested	yes
frosted	foggy	low	injured	no

snow	weather	season	physical condition	go skiing
sticky	windy	high	tired	?

43

Appendix

Bayes optimal classifier

- Suppose that we observed n samples D={x_1,…,x_n } sampled from a probability distribution P(X;θ) with a parameter θ. We want to infer how probable y is according to D.
- Method 1: Infer the parameter θ and then use P(X;θ)
  - MLE (most likely) $\theta_{MLE} = \arg\max P(D|\theta)$
  - MAP (most a posteriori) $\theta_{MAP} = \arg\max P(D|\theta)P(\theta)$
  - posterior mean

$$\hat{\theta} = \int \theta P(\theta|D) d\theta = \int \theta P(D|\theta)P(\theta)/P(D) d\theta$$

- Method 2: without inferring the parameter θ

$$P(Y,\theta|D) = P(Y,D|\theta)P(\theta)/P(D)$$

$$\longrightarrow \quad P(Y|D) = \int P(Y,D|\theta)P(\theta)/P(D) d\theta$$

Basic ideas of Bayesian inference

- Bayesian view is that we can measure uncertainty, even if there are not a lot of examples
  - What is the probability that a debut team will win the championship league this year?
    - Cannot do this with a frequentist approach
  - What is the probability that a newly minted particular coin will come up as heads?
  - Without much data we utilize our initial belief as the prior
- But as more data comes available we transfer more of our belief to the data (likelihood)
- With all the data, we do not consider the prior at all
- Belief is coded as a probability distribution

46

An example: basic ideas

- Assume that we want to infer the mean μ of a random variable x where the variance σ^2 is known and we have not yet seen any data
- $P(\mu|D,\sigma^2) = P(D|\mu,\sigma^2)P(\mu)/P(D) \propto P(D|\mu,\sigma^2)P(\mu)$
- A Bayesian would want to represent the prior μ_0 and the likelihood μ as parameterized distributions (e.g. Gaussian, multinomial, uniform, etc.)
- Let's assume a Gaussian just here
- Since the prior is a Gaussian we would like to multiply it by whatever the distribution of the likelihood is in order to get a posterior which is also a parameterized distribution specifically Gaussian

47

Conjugate Priors

- $P(\mu|D, \sigma_0^2) = P(D|\mu)P(\mu)/P(D) \propto P(D|\mu)P(\mu)$
- If the posterior is the same distribution as the prior after the multiplication, then we say the prior and posterior are *conjugate* distributions and the prior is a conjugate prior for the likelihood
- In the case of a known variance and a Gaussian prior we can use a Gaussian likelihood and the product (posterior) will also be a Gaussian
- If the likelihood is multinomial then we would need to use a Dirchlet prior and the posterior would be a Dirchlet

48

Discrete Conjugate Distributions

Discrete distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$p(\tilde{x}=1)=\frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\|\alpha',\beta')$ (beta-binomial)
Negative Binomial with known failure number r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+rn$	$\alpha-1$ total successes, $\beta-1$ failures[note 1] (i.e. $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k+\sum_{i=1}^n x_i,\ \frac{\theta}{n\theta+1}$	k total occurrences in $1/\theta$ intervals	$\operatorname{NB}(\tilde{x}\|k',\frac{\theta'}{1+\theta'})$ (negative binomial)
Poisson	λ (rate)	Gamma	$\alpha,\ \beta$ [note 3]	$\alpha+\sum_{i=1}^n x_i,\ \beta+n$	α total occurrences in β intervals	$\operatorname{NB}(\tilde{x}\|\alpha',\frac{1}{1+\beta'})$ (negative binomial)
Categorical	**p** (probability vector), k (number of categories, i.e. size of **p**)	Dirichlet	$\boldsymbol\alpha$	$\boldsymbol\alpha+(c_1,\ldots,c_k)$, where c_i is the number of observations in category i	α_i-1 occurrences of category i[note 1]	$p(\tilde{x}=i)=\frac{\alpha_i'}{\sum_i\alpha_i'}=\frac{\alpha_i+c_i}{\sum_i\alpha_i+n}$
Multinomial	**p** (probability vector), k (number of categories, i.e. size of **p**)	Dirichlet	$\boldsymbol\alpha$	$\boldsymbol\alpha+\sum_{i=1}^n \mathbf{x}_i$	α_i-1 occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\|\boldsymbol\alpha')$ (Dirichlet-multinomial)
Hypergeometric with known total population size N	M (number of target members)	Beta-binomial[4]	$n=N,\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha+n,\ \beta+\sum_{i=1}^n x_i$	$\alpha-1$ experiments, $\beta-1$ total failures[note 1]	

From Wikipedia

Continuous Conjugate Distribution (1)

Continuous distributions [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\ \sigma_0^2$	$\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)\Big/\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right),$ $\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\|\mu_0',\sigma_0^{2\prime}+\sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\ \tau_0$	$\left(\tau_0\mu_0+\tau\sum_{i=1}^n x_i\right)\Big/(\tau_0+n\tau),\ \tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\|\mu_0',\frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\ \beta$ [note 5]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu,\sigma^2=\beta'/\alpha')$[5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\ \sigma_0^2$	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\|\mu,\sigma_0^{2\prime})$[5]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\ \beta$[note 3]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu,\sigma^2=\beta'/\alpha')$[5]
Normal[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\|\mu',\frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[5]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\|\mu',\frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]

Wikipediaより

Continuous Conjugate Distribution (2)

Multivariate normal with known covariance matrix Σ	μ (mean vector)	Multivariate normal	$\boldsymbol\mu_0,\ \boldsymbol\Sigma_0$	$\left(\boldsymbol\Sigma_0^{-1}+n\boldsymbol\Sigma^{-1}\right)^{-1}\left(\boldsymbol\Sigma_0^{-1}\boldsymbol\mu_0+n\boldsymbol\Sigma^{-1}\bar{\mathbf{x}}\right),$ $\left(\boldsymbol\Sigma_0^{-1}+n\boldsymbol\Sigma^{-1}\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol\Sigma_0^{-1}$ and with sample mean $\boldsymbol\mu_0$	$\mathcal{N}(\tilde{\mathbf{x}}\|\boldsymbol\mu_0',\boldsymbol\Sigma_0'+\boldsymbol\Sigma)$[5]
Multivariate normal with known precision matrix Λ	μ (mean vector)	Multivariate normal	$\boldsymbol\mu_0,\ \boldsymbol\Lambda_0$	$\left(\boldsymbol\Lambda_0+n\boldsymbol\Lambda\right)^{-1}\left(\boldsymbol\Lambda_0\boldsymbol\mu_0+n\boldsymbol\Lambda\bar{\mathbf{x}}\right),\ \left(\boldsymbol\Lambda_0+n\boldsymbol\Lambda\right)$ • $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol\Lambda$ and with sample mean $\boldsymbol\mu_0$	$\mathcal{N}\left(\tilde{\mathbf{x}}\|\boldsymbol\mu_0',(\boldsymbol\Lambda_0'^{-1}+\boldsymbol\Lambda^{-1})^{-1}\right)$[5]
Multivariate normal with known mean μ	Σ (covariance matrix)	Inverse-Wishart	$\nu,\ \boldsymbol\Psi$	$n+\nu,\ \boldsymbol\Psi+\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol\mu)(\mathbf{x}_i-\boldsymbol\mu)^T$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\boldsymbol\Psi$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\|\boldsymbol\mu,\frac{1}{\nu'-p+1}\boldsymbol\Psi'\right)$[5]
Multivariate normal with known mean μ	Λ (precision matrix)	Wishart	$\nu,\ \mathbf{V}$	$n+\nu,\ \left(\mathbf{V}^{-1}+\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol\mu)(\mathbf{x}_i-\boldsymbol\mu)^T\right)^{-1}$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}}\|\boldsymbol\mu,\frac{1}{\nu'-p+1}\mathbf{V}'^{-1}\right)$[5]
Multivariate normal	μ (mean vector) and Σ (covariance matrix)	normal-inverse-Wishart	$\boldsymbol\mu_0,\ \kappa_0,\ \nu_0,\ \boldsymbol\Psi$	$\frac{\kappa_0\boldsymbol\mu_0+n\bar{\mathbf{x}}}{\kappa_0+n},\ \kappa_0+n,\ \nu_0+n,$ $\boldsymbol\Psi+\mathbf{C}+\frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol\mu_0)(\bar{\mathbf{x}}-\boldsymbol\mu_0)^T$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C}=\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol\mu_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol\mu_0$ and with sum of pairwise deviation products $\boldsymbol\Psi=\nu_0\boldsymbol\Sigma_0$	$t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\|\boldsymbol\mu_0',\frac{\kappa_0'+1}{\kappa_0'(\nu_0'-p+1)}\boldsymbol\Psi'\right)$[5]
Multivariate normal	μ (mean vector) and Λ (precision matrix)	normal-Wishart	$\boldsymbol\mu_0,\ \kappa_0,\ \nu_0,\ \mathbf{V}$	$\frac{\kappa_0\boldsymbol\mu_0+n\bar{\mathbf{x}}}{\kappa_0+n},\ \kappa_0+n,\ \nu_0+n,$ $\left(\mathbf{V}^{-1}+\mathbf{C}+\frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol\mu_0)(\bar{\mathbf{x}}-\boldsymbol\mu_0)^T\right)^{-1}$ • $\bar{\mathbf{x}}$ is the sample mean • $\mathbf{C}=\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol\mu_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol\mu_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu_0'-p+1}\left(\tilde{\mathbf{x}}\|\boldsymbol\mu_0',\frac{\kappa_0'+1}{\kappa_0'(\nu_0'-p+1)}\mathbf{V}'^{-1}\right)$[5]

From Wikipedia

Continuous Conjugate Distribution (3)

Uniform	$U(0,\theta)$	Pareto	$x_m,\ k$	$\max\{x_1,\ldots,x_n,x_m\},\ k+n$	k observations with maximum value x_m	
Pareto with known minimum x_m	k (shape)	Gamma	$\alpha,\ \beta$	$\alpha+n,\ \beta+\sum_{i=1}^n\ln\frac{x_i}{x_m}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)	
Weibull with known shape β	θ (scale)	Inverse gamma[4]	a, b	$a+n,\ b+\sum_{i=1}^n x_i^\beta$	a observations with sum b of the β'th power of each observation	
Log-normal with known precision τ	μ (mean)	Normal[4]	$\mu_0,\ \tau_0$	$\left(\tau_0\mu_0+\tau\sum_{i=1}^n\ln x_i\right)\Big/(\tau_0+n\tau),\ \tau_0+n\tau$	"mean" was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	
Log-normal with known mean μ	τ (precision)	Gamma[4]	$\alpha,\ \beta$[note 3]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(\ln x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance $\frac{\beta}{\alpha}$ (i.e. with sum of squared log deviations 2β — i.e. deviations between the logs of the data points and the "mean")	
Exponential	λ (rate)	Gamma	$\alpha,\ \beta$ [note 3]	$\alpha+n,\ \beta+\sum_{i=1}^n x_i$	α observations that sum to β	$\operatorname{Lomax}(\tilde{x}\|\beta',\alpha')$ (Lomax distribution)
Gamma with known shape α	β (rate)	Gamma	$\alpha_0,\ \beta_0$	$\alpha_0+n\alpha,\ \beta_0+\sum_{i=1}^n x_i$	α_0 observations with sum β_0	$\operatorname{CG}(\tilde{\mathbf{x}}\|\alpha,\alpha_0',\beta_0')=\beta'(\tilde{\mathbf{x}}\|\alpha,\alpha_0',1,\beta_0')$ [note 7]
Inverse Gamma with known shape α	β (inverse scale)	Gamma	$\alpha_0,\ \beta_0$	$\alpha_0+n\alpha,\ \beta_0+\sum_{i=1}^n\frac{1}{x_i}$	α_0 observations with sum β_0	
Gamma with known rate β	α (shape)	$\propto\frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$	$a,\ b,\ c$	$a\prod_{i=1}^n x_i,\ b+n,\ c+n$	b or c observations (b for estimating α, c for estimating β) with product a	
Gamma [4]	α (shape), β (inverse scale)	$\propto\frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r\beta^{-\alpha s}}$	$p,\ q,\ r,\ s$	$p\prod_{i=1}^n x_i,\ q+\sum_{i=1}^n x_i,\ r+n,\ s+n$	α was estimated from r observations with product p; β was estimated from s observations with sum q	

From Wikipedia