Predictive Modelling - II

k-Nearest Neighbors + Naive Bayes

Rita P. Ribeiro

Data Mining I - 2023/2024

U.PORTO
FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO
[dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

Summary

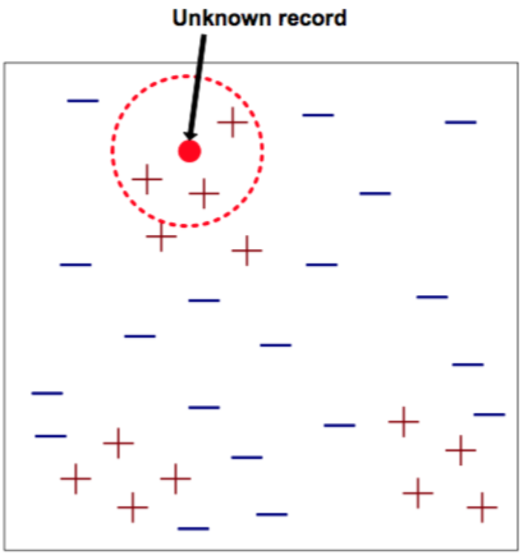
- Machine Learning: Where we at?
- Distance-based Classifiers
  - k-Nearest Neighbors}
- Probabilistic Classifiers
  - Naive Bayes
  - Bayesian Belief Networks

k-Nearest Neighbors

Predictive Modelling: Where we at?

- Distance-based Approaches
  - e.g. kNN
- Probabilistic Approaches
  - e.g. Naive Bayes, Bayesian Networks
- Mathematical Formulae
  - e.g. multiple linear regression
- Logical Approaches
  - e.g. CART
- Optimization Approaches
  - e.g. SVM, ANN
- Ensemble Approaches

k-Nearest Neighbors (kNN)



k-Nearest Neighbors (kNN)

- The k-nearest neighbor method was first described in the early 1950s.
- It is a lazy learner that does not learn any model from data, i.e. it does not learn a function to map the predictor variables into a target variable.
- It is an instance-based learning algorithm: it learns by analogy - i.e. they are based on the notion of similarity between cases.
- As it does not make any assumption on the unknown functional form we are trying to approximate, it means that with sufficient data they are applicable to any problem

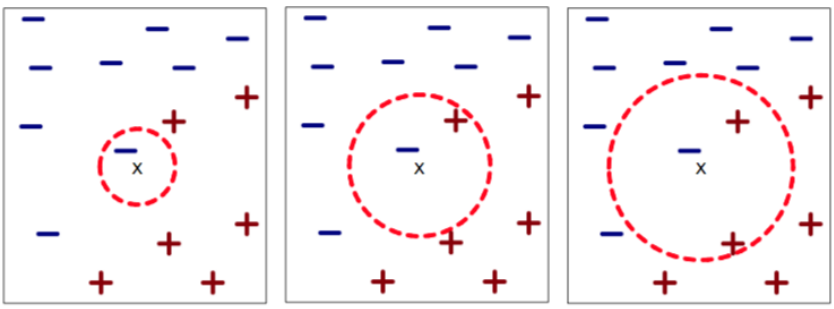
k-Nearest Neighbors (kNN)

Method:

- Choose the number *k* and the distance metric *d*
- For a test case **x**
  - find the *k* nearest cases in the training data according to *d*
  - use the target variable values of these cases to obtain the prediction for **x**
    - classification: the prediction is the majority class
    - regression: the prediction is the average of the target values

k-Nearest Neighbors (kNN)

- What should be the value of *k*?



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

k-Nearest Neighbors (kNN)

- What should be the value of k?
  - typically, 3, 5 and 7
  - odd numbers to avoid draws
  - it can be estimated experimentally:
    - global estimation searches for the ideal k for a given data set
    - local estimation methods try to estimate the ideal k for each test case (computationally very demanding!)
- What distance metrics d can be used?
  - minkowski distance (e.g. euclidean, manhattan, supremum)
  - heterogeneous distance (handling both numerical and categorical variables)

k-Nearest Neighbors (kNN): Wrap-up

- Requires a good choice of the distance metric and the value of k
  - normalization, irrelevant variables, unknown values, outliers may have a strong impact on the performance
- Frequently achieves good results
- Works well for online learning as new data is constantly arriving.
- Complexity grows linearly with the number of cases
  - needs efficient data structure implementation to search the nearest neighbors
- Fast training time, but slow testing time.

Bayesian Learning

Predictive Modelling: Where we at?

- Distance-based Approaches
  - e.g. kNN
- Probabilistic Approaches
  - e.g. Naive Bayes, Bayesian Networks
- Mathematical Formulae
  - e.g. multiple linear regression
- Logical Approaches
  - e.g. CART
- Optimization Approaches
  - e.g. SVM, ANN
- Ensemble Approaches

Motivation

Example: Disease Diagnosis

- There are two alternatives hypotheses, the patient has or does not have a certain disease.
- The probability of a patient having this disease is 8%.
- A laboratory test gives an indication of the presence (absence) of this disease:
  - it is positive (+) in 98% of the cases in which the patient has the disease;
  - it is negative (-) in 97% of the cases in which the patient does not have the disease.

Motivation

Example: Disease Diagnosis (cont.)

Disease

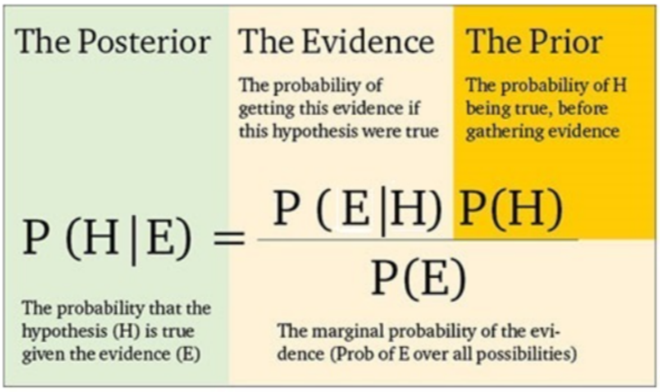
present: 8%
absent: 92%

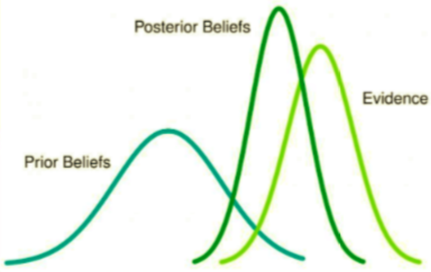
	Disease	
Lab Test	present	absent
+	98%	3%
-	2%	97%

- For a new patient the lab test is positive.
- What should the diagnosis be?
  - $P(present|+) = ?$
  - $P(absent|+) = ?$
- Bayes Theorem helps answering this question!

Motivation

Bayes Theorem





Naive Bayes

- The Naive Bayes is a particular class of Bayesian classifiers that predict the probability that a case belongs to a certain class
- It has shown rather competitive performance on several problems even when compared to more “sophisticated” methods
- It is based on the Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

where

- $P(X)$ and $P(Y)$ are the prior probabilities of X and Y, respectively
- $P(X|Y)$ is the posterior probability of X given Y
- $P(Y|X)$ is the posterior probability of Y given X

Naive Bayes

- Assume target function $f : X \rightarrow Y$, where each instance x described by p attributes $\langle x_1, x_2 \ldots x_p \rangle$.
- Most probable value of $f(x)$ is:

$$
\begin{aligned}
\hat{y} &= \operatorname*{argmax}_{y_j \in Y} P(y_j | x_1, x_2 \ldots x_p) \\
\hat{y} &= \operatorname*{argmax}_{y_j \in Y} \frac{P(x_1, x_2 \ldots x_p | y_j) P(y_j)}{P(x_1, x_2 \ldots x_p)} \\
&= \operatorname*{argmax}_{y_j \in Y} P(x_1, x_2 \ldots x_p | y_j) P(y_j)
\end{aligned}
$$

Naive Bayes

Naive Bayes Assumption

Attributes are independent given the class.

$$P(x_1, x_2 \ldots x_n | y_j) = \prod_i P(x_i | y_j)$$

which gives

$$\hat{y} = \operatorname*{argmax}_{y_j \in V} P(y_j) \prod_i P(x_i | y_j)$$

Naive Bayes

Example: Disease Diagnosis (cont.)

- there is only one attribute
  - the lab test outcome: positive (+) or negative (-)

Disease

present: 8%
absent: 92%

	Disease	
Lab Test	present	absent
+	98%	3%
-	2%	97%

- for a new patient, the lab test is positive (+)

- what should be the diagnose?
  - $P(present|+) \propto P(+|present) \times P(present) = 0.98 \times 0.08 = 0.0784$
  - $P(absent|+) \propto P(+|absent) \times P(absent) = 0.03 \times 0.92 = 0.0276$

Naive Bayes

How to estimate the probabilities?

- Assume a decision problem with p predictor variables.
- Each variable assume k values.
- The joint probability requires to estimate k^p probabilities.
- Assuming that variables are conditionally independent given the class, only requires to estimate $k \times p$ probabilities.
- For categorical attributes, the probability is estimated from frequency tables.
- But, how to do for numeric attributes?

Naive Bayes

How to estimate the probabilities? (cont.)

- Gaussian Naive Bayes
  - For a given value x_k of a numeric attribute X, we estimate the probability given a class y_j, assuming a Normal distribution, i.e.

$$P(x_k|y_j) = \frac{1}{\sigma_{kj}\sqrt{2\pi}} e^{-\frac{(x_k-\mu_{kj})^2}{2\sigma_{kj}^2}}$$

    where μ_{kj} and σ_{kj} are the mean and the standard deviation of the values of attribute X for which the class is y_j.
- Other solutions exist.

Naive Bayes: Example

Example: Is it good to play golf?

Weather	Temperature	Humidity	Wind	Play
Rainy	71	91	Yes	No
Sunny	69	70	No	Yes
Sunny	80	90	Yes	No
Overcast	83	86	No	Yes
Rainy	70	96	No	Yes
Rainy	65	70	Yes	No
Overcast	64	65	Yes	Yes
Overcast	72	90	Yes	Yes
Sunny	75	70	Yes	Yes
Rainy	68	80	No	Yes
Overcast	81	75	No	Yes
Sunny	85	85	No	No
Sunny	72	95	No	No
Rainy	75	80	No	Yes

Naive Bayes: Example

Example: Is it good to play golf?

- Estimate probabilities from data
- Nr. examples: 14, $P(Play =' Yes') = 9/14$, $P(Play =' No') = 5/14$

Weather			Temperature			Humidity			Wind		
	Yes	No		Yes	No		Yes	No		Yes	No
Sunny	2/9	3/5	μ	73	74.6	μ	79.1	86.2	False	6/9	2/5
Overcast	4/9	0/5	σ	6.2	7.9	σ	10.2	9.3	True	3/9	3/5
Rainy	3/9	2/5									

- Estimated probability for a value of temperature
  - $P(Temp = 66|\, Yes) = \frac{1}{6.2\sqrt{2\pi}} e^{-\frac{(66-73)^2}{2(6.2)^2}} = 0.03401871$

Naive Bayes: Example

Example: Is it good to play golf?

Weather	Temperature	Humidity	Wind	Play
Sunny	66	90	Yes	?

$$P(Yes|\,Weather = Sunny, Temperature = 66, Humidity = 90, Wind = Yes) =$$
$$P(Yes)P(Weather = Sunny|\,Yes)P(Temperature = 66|\,Yes)P(Humidity = 90|\,Yes)P(Wind = Yes|\,Yes)$$
$$\approx 0.000028$$

$$P(No|\,Weather = Sunny, Temperature = 66, Humidity = 90, Wind = Yes) =$$
$$P(No)P(Weather = Sunny|No)P(Temperature = 66|No)P(Humidity = 90|No)P(Wind = Yes|No)$$
$$\approx 0.000015$$

Prediction: Play = Yes

Naive Bayes: Laplace Correction

- If one of the conditional probabilities is equal to zero, the entire expression becomes zero
- Use other estimates of conditional probabilities
- Laplace: $P(x_i|y_j) = \frac{n_{ij}+1}{n_j+m}$, where
    - n_{ij} is nr. examples for which $Y = y_j$ and $X = x_i$
    - n_j is nr. examples for which $Y = y_j$
    - m is a weight greater than zero (typically 1)
- There are other corrections . . .

Naive Bayes: Summary

- The variability of a dataset is summarized in contingency tables.
- Robust to noise or irrelevant values that do not have a strong statistical support; but, redundant variables can be a problem.
- Requires a single scan over the dataset.
- The algorithm is Incremental (incorporation of new examples) and decremental (forgetting old examples).
- The dimension of the decision model is independent of the number of examples.
- If the independence assumption does not hold for some attributes
    - use other techniques (e.g. Bayesian (Belief) Networks)

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Ferreira, Pedro G. 2019. “Fundamentals and Applications of ML Course.” Slides.

Flach, Peter. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511973000.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining*. 2nd ed. Pearson.

Torgo, Luís. 2017. “Data Mining i Course.” Slides.