Classification

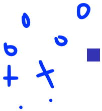
KNN Classifier, Naive Bayesian Classifier

Outline

k-Nearest-Neighbor classifiers

- First introduced in early 1950s
- Distance metrics
  - Generally Euclidean distance is used when the values are continuous
  - Hamming distance for categorical/nominal values

Algorithm idea



- Let k be the number of nearest neighbors and D be the set of training examples
- **for** each test example $z = (\mathbf{x}', y')$, do
  - Compute $d(\mathbf{x}, \mathbf{x}')$, the distance between z and every example $(\mathbf{x}, y) \in D$
  - Select $D_z \subseteq D$, the set of k closest training examples to z
  - $y' = argmax_v \sum_{(x_i, y_i) \in D_z} I(v == y_i)$ //Majority voting
- **end for**

v: a class label

y_i: the class label for one of the NNs

I: indicate function that returns value 1 if its argument is true and 0 otherwise.

Discussions

- Computation can be costly if the number of training examples is large.
  - Efficient indexing techniques are available to find the nearest neighbors of a test example.
- Sensitive k: reduce the impact of k is to weight the influence of a NN $\mathbf{x}_i$: $w_i = \frac{1}{d(\mathbf{x}', \mathbf{x}_i)^2}$
  Distance-weighted voting:

$$y' = argmax_v \sum_{(x_i, y_i) \in D_z} w_i \times I(v == y_i)$$

Characteristics

- Instance-based learning: (1) require a proximity measure (2) make predictions without having to maintain a model
- Lazy learner: do not require model building.
- Predictions are made based on local information. Thus, it is quite susceptible to noise.
- Can produce arbitrarily shaped decision boundaries. More flexible compared with the decision tree approach.

References

- Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, Eamonn J. Keogh: Querying and mining of time series data: experimental comparison of representations and distance measures. PVLDB 1(2):1542-1552 (2008)
- Abdullah Mueen, Eamonn J. Keogh, Neal Young: Logical-shapelets: an expressive primitive for time series classification. KDD 2011:1154-1162.
- Minh Nhut Nguyen, Xiao-Li Li, See-Kiong Ng: Positive Unlabeled Learning for Time Series Classification. IJCAI2011: 1421-1426.

KNeighborsClassifier – Python scikit-learn 0.22.1

```
classsklearn.neighbors.KNeighborsClassifier(
    n_neighbors=5,weights='uniform',algorithm='auto',
    leaf_size=30,p=2,metric='minkowski',
    metric_params=None,n_jobs=None,**kwargs)
```

- **n_neighbors**: Number of neighbors to use by default for kneighbors queries.
- **p**: Power parameter for the Minkowski metric. When p = 1, this is equivalent to using manhattan_distance (l1), and euclidean_distance (l2) for p = 2. For arbitrary p, minkowski_distance (l_p) is used.
- **Metric**: the distance metric to use. The default metric is minkowski, and with p=2 is equivalent to the standard Euclidean metric.

Naive Bayesian Classifier

- Non-deterministic relationship between attribute set and the class label.
  - Reason 1: noisy data
  - Reason 2: complex factors
- Probabilistic relationships between attribute set and the class label.
- Bayesian classifiers: a probabilistic framework for solving classification problems.

Bayes Theorem

- Conditional probability

$$P(C|A) = \frac{P(A, C)}{P(A)}$$

$$P(A|C) = \frac{P(A, C)}{P(C)}$$

- Bayes theorem

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?
- Solve this problem:
  - M: a patient has meningitis
  - S: a patient has stiff neck
  - $\underline{P(S|M) = 50\%}$, $\underline{P(M) =}\ 1/50,000$, $\underline{P(S) =}\ 1/20$, $P(M|S)$=?

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{(1/2)\cdot(1/50,000)}{1/20} = \frac{20}{100,000} = 0.0002$$

Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes $(A_1, A_2, \cdots, A_n)$
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes $P(C|A_1, A_2, \cdots, A_n)$
- Can we estimate $P(C|A_1, A_2, \cdots, A_n)$ directly from data?

Bayesian Classifiers – approach

- Compute the posterior probability $P(C|A_1, A_2, \cdots, A_n)$ for all values of C using the Bayes theorem

$$P(C|A_1, A_2, \cdots, A_n) = \frac{P(A_1, A_2, \cdots, A_n|C)P(C)}{P(A_1, A_2, \cdots, A_n)}$$

- Choose value of C that maximizes $P(C|A_1, A_2, \cdots, A_n)$

$$\equiv maximize \; P(A_1, A_2, \cdots, A_n|C)P(C)$$

- How to estimate $P(A_1, A_2, \cdots, A_n|C)$?

Naive Bayes Classifier

P(A1, A2) = P(A1) P(A2)
P(A1, A2) = P(A1) P(A2|A1)

- Assume independence among attributes A_i when class is given

$$P(A_1, A_2, \cdots, A_n|C_j) = P(A_1|C_j)P(A_2|C_j)\cdots P(A_n|C_j)$$

- Can estimate $P(A_i|C_j)$ for all A_i and C_j.
- New point is classified to C_j if $P(C_j)\Pi P(A_i|C_j)$ is maximal.

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(C)
P(C=Yes) = 3/10
P(C=No) = 7/10

- Class: $P(C = C_i) = \frac{N_i}{N}$
- Discrete attributes: $P(A_i = a_i | C = C_j) = \frac{N_{ij}}{N_j}$
  - N_{ij}: # of instances having attribute $A_i = a_i$ and belonging to class C_j
  - N_j: # of instances belonging to class C_j
- P(No) $= 0.7$, P(Yes) $= 0.3$
- $P(Status = Married | No) = \frac{4}{7}$ $P(Refund = Yes | Yes) = 0$

Estimate Probabilities from Data – continuous attributes

- Discretize the range into bins
  - one ordinal attribute per bin
- Two-way split: $(A < v)$ or $(A \geq v)$
  - choose only one of the two splits as new attribute
- Probability density estimation
  - Assume attribute follows a normal distribution
  - Use data to estimate parameters of distribution (e.g., mean μ and standard deviation σ)
  - Once probability distribution is known, can use it to estimate the conditional probability

$$P(A_i|c_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - μ_{ij}: estimated as sample mean of A_i for all training records with class c_j.
  - σ_{ij}: estimated as sample variance s^2 of all training records with class c_j.

Estimate Probabilities from Data – Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

For (Income, Class=No)
Sample mean: $\frac{125+100+70+120+60+220+75}{7} = 110$
Sample variance: 2975

$$P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} exp^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naive Bayes Classifier – Example 1

A1

Given a test record: X=(Refund=No, ~~Married~~, Income=120K)

P(C|A1=No, A2=M, A3=120K)
~ P(A1=No|C) P(A2=M|C) P(A3=120K|C) P(C)

$P(Refund = Yes|No) = 3/7$
$P(Refund = No|No) = 4/7$
$P(Refund = Yes|Yes) = 0$
$P(Marital\ Status = Single|No) = 2/7$
$P(Marital\ Status = Divorced|No) = 1/7$
$P(Marital\ Status = Married|No) = 4/7$
$P(Marital\ Status = Single|Yes) = 2/3$
$P(Marital\ Status = Divorced|Yes) = 1/3$
$P(Marital\ Status = Married|Yes) = 0$
$P(Class = Yes) = 0.3$
$P(Class = No) = 0.7$

P(A1=No|C=Y) P(A2=M|C=Y) P(A3=120K|C=Y) P(C=Y)

>

P(A1=No|C=N) P(A2=M|C=N) P(A3=120K|C=N) P(C=N)

For taxable income

If class=No: Sample mean=110, variance=2975

If class=Yes: Sample mean=90, variance=25

Example 1 (cont.)

$$
\begin{aligned}
P(X|Class = No) =& P(Refund = No|No) * P(Married|No) * P(Income = 120K|No) \\
=& 4/7 * 4/7 * 0.0072 \\
=& 0.0024
\end{aligned}
$$

$$
\begin{aligned}
P(X|Class = Yes) =& P(Refund = No|Yes) * P(Married|Yes) * P(Income = 120K|Yes) \\
=& 1 * 0 * 1.2 * 10^{-9} \\
=& 0
\end{aligned}
$$

Since $\underline{P(X|No) * P(No)} > \underline{P(X|Yes) * P(Yes)}$
$P(No|X) > P(Yes|X) \rightarrow$ class = No

$P(Married|Yes) = 0$. If one of the conditional probability is zero, then the entire expression becomes zero.

Naive Bayes Classifier – Probability estimation for Zero conditional probability

- Original: $P(A_i = a_i | C = C_j) = \frac{N_{ij}}{N_j}$
- Laplace: $P(A_i = a_i | C = C_j) = \frac{N_{ij}+1}{N_j+c}$
- m-estimate: $P(A_i = a_i | C = C_j) = \frac{N_{ij}+mp}{N_j+m}$
- N_{ij}: the number of instances with attribute value a_i and belonging to a class C_j
- N_j: the number of instances belonging to a class C_j
- c: total number of classes
- p: prior probability for the positive class C_j
- m: parameter

E.g., use m-estimate: $m = 3$ and $p = 0.3$ (prior for positive class *Yes*)
P(Marital Status=Married|Yes)$=\frac{0+0.3*3}{3+3} = \frac{0.9}{6}$

Example of Naive Bayes Classifier – Example 2

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Example 2 (cont.)

Give test case:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
	yes	no	yes	no	?

- A: attributes
- M: mammals
- N: non-mammals

Class M: P(GB=Yes|M) P(CF=No|M) P(LW=yes|M) P(HL=no|M). P(M)

Class N: P(GB=Yes|N) P(CF=No|N) P(LW=yes|N) P(HL=no|N). P(N)

Example 2 (cont.)

- $\mathrm{P(M)} = \frac{7}{20}$, $\mathrm{P(N)} = \frac{13}{20}$
- $P(GB = Y|M) = \frac{6}{7}$, $P(GB = N|M) = \frac{1}{7}$
- $P(Can\ Fly = Y|M) = \frac{1}{7}$, $P(Can\ Fly = N|M) = \frac{6}{7}$
- $P(Live\ in\ Water = Y|M) = \frac{2}{7}$, $P(Live\ in\ Water = N|M) = \frac{5}{7}$
- $P(Have\ Legs = Y|M) = \frac{5}{7}$, $P(Have\ Legs = N|M) = \frac{2}{7}$

$$
\begin{aligned}
P(X|M) =& P(GB = Y|M) * P(Can\ Fly = N|M) * P(Live\ in\ Water = Y|M) \\
& * P(Have\ Legs = N|M) \\
=& 6/7 * 6/7 * 2/7 * 2/7 = 0.06
\end{aligned}
$$

$$
\begin{aligned}
P(X|N) =& P(GB = Y|N) * P(Can\ Fly = N|N) * P(Live\ in\ Water = Y|N) \\
& * P(Have\ Legs = N|N) \\
=& 1/13 * 10/13 * 3/13 * 4/13 = 0.0042
\end{aligned}
$$

$P(X|M) * P(M) = 0.06 * 7/20 = 0.021$
$P(X|N) * P(N) = 0.0042 * 13/20 = 0.0027$

Since $P(X|M) * P(M) > P(X|N) * P(N)$, class = Mammals

Naive Bayes Classifier (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
    - Use other techniques such as Bayesian Belief Networks (BBN)

References

- Chapter 4: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar
- KNN classifier: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- Naive Bayes classifier: https://scikit-learn.org/stable/modules/naive_bayes.html