Chapter 9 Evidence and Probabilities

1/3/21 5:09 PM

*A broadly applicable framework both for **evaluating the evidence**, and for combining it to **estimate the resulting likelihood***

Combining Evidence Probabilistically

Bayes Rule for classification

$$p(C = c \mid \mathbf{E}) = \frac{p(\mathbf{E} \mid C = c) \cdot p(C = c)}{p(\mathbf{E})}$$

1. Left hand side: probability of target variable C taking class of interest c, after taking the evidence E into account
2. P(C=c) is the 'prior' probability of the class, could come from several places:
   a. "subjective prior", meaning it is a belief of particular decision maker based on her knowledge and experience
   b. 'prior' belief based on previous applications of Bayes Rule with other evidence
   c. Unconditional probability inferred from data, usually from class prior / base rate of c, i.e. percentage of all examples in data that are of class c
3. p(E |C = c) is the likelihood of seeing evidence E, given C=c.
   a. Can be calculated from the percentage of examples of class c, which have feature vector of E
4. P(E) is likelihood of evidence E in dataset
   a. Can be calculated from percentage occurrence of E from among all examples

Conditional Independence and Naïve Bayes

Assume attributes are conditional independent, given the class
(So each evidence class can be decomposed)

Bayes Equation:

$$p(c \mid \mathbf{E}) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdots p(e_k \mid c) \cdot p(c)}{p(\mathbf{E})}$$

Naïve Bayes Classifier: estimate the probability that the example belongs to each class and reports the class with highest probability.

Advantages and Disadvantages of Naïve Bayes

1. Efficient in terms of storage space and computation time
   - Training consists of only storage of counts of classes and feature occurrences
2. Violation of independence assumption tend not to hurt classification performance.
   - E.g. double-counting two strongly correlated evidence would result in a more extreme classifier, but as long as the evidence shows right direction, the classifier would still work well
   - But need to be careful when we care the actual value of probability
     E.g. when combining with cost-benefits
3. Incremental learner
   - Update model one training example at a time. Doesn't need to re-process all previous training examples when new training data becomes available

A Model of Evidence Lift

Lift

- measures how much more prevalent the positive class is in the selected subpopulation over the prevalence in the population as a whole

$$\mathrm{lift}_c(x) = \frac{p(x \mid c)}{p(x)}$$

If we assume full feature independence (Naïve-Naïve Bayes)

$$p(c \mid \mathbf{E}) = \frac{p(e_1 \mid c) \cdot p(e_2 \mid c) \cdots p(e_k \mid c) \cdot p(c)}{p(e_1) \cdot p(e_2) \cdots p(e_k)}$$

$$p(C = c \mid \mathbf{E}) = p(C = c) \cdot \mathrm{lift}_c(e_1) \cdot \mathrm{lift}_c(e_2) \cdots$$

An Example from Facebook "Likes"

Like	Lift	Like	Lift
Lord Of The Rings	1.69	Wikileaks	1.59
One Manga	1.57	Beethoven	1.52
Science	1.49	NPR	1.48
Psychology	1.46	*Spirited Away*	1.45
The Big Bang Theory	1.43	Running	1.41
Paulo Coelho	1.41	Roger Federer	1.40
The Daily Show	1.40	*Star Trek*	1.39
Lost	1.39	Philosophy	1.38
Lie to Me	1.37	*The Onion*	1.37
How I Met Your Mother	1.35	*The Colbert Report*	1.35
Doctor Who	1.34	*Star Trek*	1.32
Howl's Moving Castle	1.31	Sheldon Cooper	1.30
Tron	1.28	*Fight Club*	1.26
Angry Birds	1.25	*Inception*	1.25
The Godfather	1.23	*Weeds*	1.22

- The lift for High-IQ class
- (Proportion of "likes" among High-IQ class) / (Proportion of "likes" among all users)
- The base rate of High-IQ class is 0.14. If I give "like" to nothing, then probability of being high-IQ is just 0.14
- If I give like to Sheldon Cooper, then probability of High-IQ increases to 0.14 * 1.3 = 0.182
- If I have 3 likes, Sheldon Cooper, Star Trek and Lord of Rings, then my probability of being High-IQ is 0.14 * 1.3 * 1.39 * 1.69 = 0.4275