Practice 4: Bayes' Theorem

Seong-Ho Choi

Seoul National University

shchoi@bi.snu.ac.kr

May 30, 2017

Overview

Bayes' Theorem

Bayes' Theorem: Suppose that E and F are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then:

$$p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F) + p(E|\bar{F})p(\bar{F})}$$

Example: We have two boxes. The first box contains two green balls and seven red balls. The second contains four green balls and three red balls. Bob selects one of the boxes at random. Then he selects a ball from that box at random. If he has a red ball, what is the probability that he selected a ball from the first box.

- Let E be the event that Bob has chosen a red ball and F be the event that Bob has chosen the first box.
- By Bayes' theorem the probability that Bob has picked the first box is:

$$p(F|E) = \frac{(7/9)(1/2)}{(7/9)(1/2) + (3/7)(1/2)} = \frac{7/18}{38/63} = \frac{49}{76} \approx 0.645$$

Derivation of Bayes' Theorem

- Recall the definition of the conditional probability $p(E|F)$:

$$p(E|F) = \frac{p(E \cap F)}{p(F)}$$

- From this definition, it follows that:

$$p(E|F) = \frac{p(E \cap F)}{p(F)}, \quad p(F|E) = \frac{p(E \cap F)}{p(E)}$$

$$p(E|F)p(F) = p(E \cap F), \quad p(F|E)p(E) = p(E \cap F)$$

- Equating the two formulas for $p(E \cap F)$ shows that

$$p(E|F)p(F) = p(F|E)p(E)$$

- Solving for $p(E|F)$ and for $p(F|E)$ tells us that

$$p(E|F) = \frac{p(F|E)p(E)}{p(F)}, \quad p(F|E) = \frac{p(E|F)p(F)}{p(E)}$$

Derivation of Bayes' Theorem

- On the last slide we showed that

$$p(F|E) = \frac{p(E|F)p(F)}{p(E)}$$

- Note that $p(E) = p(E|F)p(F) + p(E|\bar{F})p(\bar{F})$

$$since \quad p(E) = p(E \cap F) + p(E \cap \bar{F}) = p(E|F)p(F) + p(E|\bar{F})p(\bar{F})$$

- Hence,

$$p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F) + p(E|\bar{F})p(\bar{F})}$$

Genrealized Bayes' Theorem

Generalized Bayes' Theorem Suppose that E is an event from a sample space S and that F_1, F_2, ..., F_n are mutually exclusive events such that $\bigcup_{i}^{n} F_i = S$.

Assume that $p(E) \neq 0$ for $i = 1, 2, ..., n$. Then

$$p(F_j|E) = \frac{p(E|F_j)p(F_j)}{\sum_{i=1}^{n} p(E|F_i)p(F_i)}$$

Applying Bayes' Theorem

Example: Suppose that one person in 100,000 has a particular disease. There is a test for the disease that gives a positive result 99% of the time when given to someone with the disease. When given to someone without the disease, 99.5% of the time it gives a negative result. Find

1. the probability that a person who test positive has the disease.
2. the probability that a person who test negative does not have the disease.

- Let D be the event that the person has the disease, and E be the event that this person tests positive.
- $p(D) = 0.000001$
- $p(E|D) = 0.99$
- $p(\bar{E}|\bar{D}) = 0.995$

Applying Bayes' Theorem

1. the probability that a person who test positive has the disease.

$$p(D|E) = \frac{p(E|D)p(D)}{p(E|D)p(D) + p(E|\bar{D})p(\bar{D})}$$
$$= \frac{(0.99)(0.00001)}{(0.99)(0.00001) + (0.005)(0.99999)} \approx 0.002$$

2. the probability that a person who test negative does not have the disease.

$$p(\bar{D}|\bar{E}) = \frac{p(\bar{E}|\bar{D})p(\bar{D})}{p(\bar{E}|\bar{D})p(\bar{D}) + p(\bar{E}|D)p(D)}$$
$$= \frac{(0.995)(0.99999)}{(0.995)(0.99999) + (0.01)(0.00001)} \approx 0.9999999$$

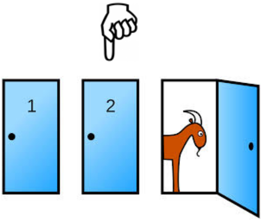
Applying Bayes' Theorem

(HW#5) **10**: An electronics company is planning to introduce a new camera phone. The company commissions a marketing report for each new product that predicts either the success or the failure of the product. Of new products introduces by the company, 60% have been successes. Furthermore, 70% of their successful products were predicted to be successes, while 40% of failed products were predicted to be successes. Find the probability this new camera phone will be successful if its success has been predicted.

- Let E be the event that the product success has been predicted, and F be the event that the product will be successful.

$$p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F) + p(E|\bar{F})p(\bar{F})} = \frac{(0.7)(0.6)}{(0.7)(0.6) + (0.4)(0.4)}$$

Monty Hall Puzzle



You are asked to select one of the three doors to open. There is a large prize behind one of the doors and if you select that door, you win the prize. After you select a door, the game show host opens one of the other doors (which he knows is not the winning door). Should you switch?

Approach 1) Because the host always opens one of the wrong door, you will surely win the prize by swithching when you were wrong at the fist time, which is $2/3$.

Monty Hall Puzzle

Approach 2) By the Bayes' rule, the conditional probability of winning by switching is

$$P(A_x|B_y, C_z) = \frac{P(B_y|A_x, C_z)P(A_x|C_z)}{P(B_y|C_z)},$$

where A_x is the event space that the car is behind door x, B_y is the space of the host's choice door y, and C_z is that of your first pick, door z. (knowing that it works on arbitrary x, y, z)

$$= \frac{P(B_y|A_x, C_z)P(A_x|C_z)}{P(B_y|A_x, C_z)P(A_x|C_z) + P(B_y|A_y, C_z)P(A_y|C_z) + P(B_y|A_z, C_z)P(A_z|C_z}$$

$$= \frac{P(B_y|A_x, C_z)}{P(B_y|A_x, C_z) + P(B_y|A_y, C_z) + P(B_y|A_z, C_z)}$$

$$= \frac{1}{1 + 0 + 1/2}$$

$$= \frac{2}{3}$$

Bayesian Spam Filters

- Let S be the event that the message is spam, and E be the event that the message conatains the word w.
- Using Bayes' Rule,

$$p(S|E) = \frac{p(E|S)p(S)}{p(E|S)p(S) + p(E|\bar{S})p(\bar{S})}$$

- Assuming that it is equally likely that an arbitrary message is spam and is not spam; i.e., $p(S) = 1/2$.

$$p(S|E) = \frac{p(E|S)}{p(E|S) + p(E|\bar{S})}$$

- Using our empirical estimates of $p(E|S)$ and $p(E|\bar{S})$.

$$r(w) = \frac{p(w)}{p(w) + q(w)}$$

- $r(w)$ estimates the probability that the message is spam.

Bayesian Spam Filters

(p.50) **Example**: We find that the word "Rolex" occurs in 250 out of 2000 spam messages and occurs in 5 out of 1000 non-spam messages. Estimate the probability that an incoming message is spam. Suppose our threshold for rejecting the email is 0.9.

$$r(Rolex) = \frac{p(Rolex)}{p(Rolex) + q(Rolex)} = \frac{0.125}{0.125 + 0.005} \approx 0.962$$

Bayesian Spam Filters using Multiple Words

- Consider the case where E_1 and E_2 denote the events that the message contains the words w_1 and w_2 respectively.

$$p(S|E_1 \cap E_2) = \frac{p(E_1|S)p(E_2|S)}{p(E_1|S)p(E_2|S) + p(E_1|\bar{S})p(E_2|\bar{S})}$$

$$r(w_1, w_2) = \frac{p(w_1)p(w_2)}{p(w_1)p(w_2) + q(w_1)q(w_2)}$$

- In general, the more words we consider, the more accurate the spam filter. With the independence assumption if we consider k words:

$$p(S|\bigcap_{i=1}^{k} E_i) = \frac{\prod_{i=1}^{k} p(E_i|S)}{\prod_{i=1}^{k} p(E_i|S) + \prod_{i=1}^{k} p(E_i|\overline{S})}$$

$$r(w_1, ..., w_n) = \frac{\prod_{i=1}^{k} p(w_i)}{\prod_{i=1}^{k} p(w_i) + \prod_{i=1}^{k} q(w_i)}$$

(p.52) **Example:** We have 2000 spam messages and 1000 non-spam messages. The word "stock" occurs 400 times in the spam messages and 60 times in the non-spam. The word "undervalued" occurs in 200 spam messages and 25 non-spam.

$$r(stock, undervalued) = \frac{p(stock)p(undervalued)}{p(stock)p(undervalued) + q(stock)q(undervalued)}$$
$$= \frac{(0.2)(0.1)}{(0.2)(0.1) + (0.06)(0.025)} \approx 0.930$$

Bayesian Spam Filters using Multiple Words

(2016-1 final) 9. Suppose that a Bayesian spam filter is trained on a set of 10,000 spam messages and 5000 messages that are not spam. The word "enhancement" appears in 1500 spam messages and 20 messages that are not spam, while the word "herbal" appears in 800 spam messages and 200 messages that are not spam. Estimate the probability that a received message containing both the words "enhancement" and "herbal" is spam.

$$r(enhancement, herbal)$$

$$= \frac{p(enhancement)p(herbal)}{p(enhancement)p(herbal) + q(enhancement)q(herbal)}$$

$$= \frac{(0.3)(0.16)}{(0.3)(0.16) + (0.004)(0.4)} \approx 0.968$$