

Name: Elise McElhiney

1. 2.4 Bayes rule for medical diagnosis

(Source: Koller.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

Solution: We want to find the probability that we have the rare disease given that we test positive.

$$\text{Bayes rule: } P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

In this case:

$$P(\text{disease}) = 0.0001$$

$$P(\text{positive}|\text{disease}) = 0.99$$

$$P(\text{positive}|\overline{\text{disease}}) = 0.01$$

$$P(\text{positive}) = P(\text{positive}|\text{disease})P(\text{disease}) + P(\text{positive}|\overline{\text{disease}})P(\overline{\text{disease}})$$

so, by Bayes rule

$$P(\text{disease}|\text{positive}) = \frac{P(\text{disease})P(\text{positive}|\text{disease})}{P(\text{positive})} = \frac{0.0001 * 0.99}{0.99 * 0.0001 + 0.01 * 0.999} \approx 0.0098$$

2. 2.5 The Monty Hall Problem

(Source: Mackay.) On a game show, a contestant is told the rules as follows:

There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of the doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially the prize is equally likely to be behind any of the 3 doors.

Solution: (From taking information theory) The logical solution to this is to realize that if you pick an incorrect door then Monty gives you the correct answer by revealing the following incorrect door. Therefore since the chances of picking the correct door on the first attempt are $P(\text{correct}) = \frac{1}{3}$ and therefore $P(\text{incorrect}) = \frac{2}{3}$. If we select an incorrect door on our first attempt and change to the other option, we win with the same probability as picking an incorrect door. Therefore if we change doors on our second chance, $P(\text{win}) = \frac{2}{3}$, where if we stay with our initial selection $P(\text{win}) = \frac{1}{3}$

We want to find the probability that the prize is behind door 2 given that Monty opens door 3

$$\text{Bayes rule: } P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Initially we know that the prize will be behind, our selected door, door 1 with a $P(1) = \frac{1}{3}$

$$P(2|3\text{opened}) = \frac{P(2)P(3\text{opened}|2)}{P(3\text{opened})} = \frac{\frac{1}{3} * 1}{\frac{1}{2} * \frac{1}{3} + 1 * \frac{1}{3}} = \frac{2}{3}$$

If we stick with our original selection $P(\text{win}) = \frac{1}{3}$, however if we change, $P(\text{win}) = \frac{2}{3}$ so we should (b) change to the other option.

3. 2.6 Conditional Independence

(Source: Koller.)

a. Let $H \in \{1, \dots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2 . Suppose we wish to calculate the vector:

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), \dots, P(H = K|e_1, e_2))$$

Which of the following numbers are sufficient for the calculation?

i. $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$

ii. $P(e_1, e_2), P(H), P(e_1, e_2|H)$

iii. $P(e_1|H), P(e_2|H), P(H)$

b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H)

Which of the above 3 sets are sufficient now?

Solution:

a.

i. is insufficient

ii. is sufficient since: $P(H|e_1, e_2) = \frac{P(H)P(e_1, e_2|H)}{P(e_1, e_2)}$

iii. is insufficient

b.

Given conditional independence:

$$P(e_1, e_2|H) = P(e_1|H) * P(e_2|H)$$

$$P(e_1, e_2) = P(e_1, e_2|H)P(H) = P(e_1|H) * P(e_2|H) * P(H)$$

i. is sufficient since: $P(H|e_1, e_2) = \frac{P(H)(P(e_1|H)*P(e_2|H))}{P(e_1, e_2)}$

ii. is sufficient since: $P(H|e_1, e_2) = \frac{P(H)P(e_1, e_2|H)}{P(e_1, e_2)}$

iii. is sufficient since: $P(H|e_1, e_2) = \frac{P(H)(P(e_1|H)*P(e_2|H))}{P(e_1|H)*P(e_2|H)*P(H)}$

4. 3.19 Irrelevant features with naive Bayes

a. Assuming $p(C = 1) = p(C = 2) = 0.5$, write down an expression for the log posterior odds ratio,

Solution:

a.

$$\log_2 p(C|x_i) = \log_2 p(x_i|C)p(C) = \log_2 p(x_i|C) + \log_2 p(C) = \varphi(x_i)^T \beta_c + \log_2 p(C)$$

$$\begin{aligned} \log_2 \frac{p(c=1|x_i)}{p(c=2|x_i)} &= \log_2 p(c=1|x_i) - \log_2 p(c=2|x_i) \\ &= \varphi(x_i)^T \beta_1 + \log_2 p(c=1) - \varphi(x_i)^T \beta_2 - \log_2 p(c=2) \\ &= \varphi(x_i)^T \beta_1 + \log_2 0.5 - \varphi(x_i)^T \beta_2 - \log_2 0.5 \\ &= \varphi(x_i)^T \beta_1 - \varphi(x_i)^T \beta_2 \\ &= \varphi(x_i)^T (\beta_1 - \beta_2) \end{aligned}$$

b.

A word will have no effect on the class posterior if $\beta_{1,w} = \beta_{2,w}$

c.

This word will not be ignored since we know $n_1 \neq n_2$ and as such:

$$\hat{\theta}_{c,w} = \frac{1 + \sum_{i \in c} x_{i,w}}{2 + n_c} = \frac{1 + n_c}{2 + n_c}$$

Since if $n_1 \neq n_2$ we know $\frac{1+n_1}{2+n_1} \neq \frac{1+n_2}{2+n_2}$ and we also know $\hat{\theta}_{1,w} \neq \hat{\theta}_{2,w}$ and as such, the word will not be ignored.

d.

We could just ignore words that occur in too high a percentage of all documents. This would prevent words like "and", "the", "or" from becoming impactful classifiers even though they are so frequent as to be irrelevant.

5. 3.22 Fitting a naive Bayes spam filter by hand

Vocabulary = "secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza"

Spam = "million dollar offer", "secret offer today", "secret is secret"

Non-spam = "low price for valued customer", "play secret sports today", "sports is healthy", "low price pizza"

Solution:

$$\hat{\theta}(\text{spam}) = \frac{\text{spam}}{\text{total}} = \frac{3}{7}$$

$$\hat{\theta}(\text{secret}|\text{spam}) = \frac{\text{secret}|\text{spam}}{\text{spam}} = \frac{2}{3}$$

$$\hat{\theta}(\text{secret}|\text{non-spam}) = \frac{\text{secret}|\text{non-spam}}{\text{non-spam}} = \frac{1}{4}$$

$$\hat{\theta}(\text{sports}|\text{non-spam}) = \frac{\text{sports}|\text{non-spam}}{\text{non-spam}} = \frac{2}{4}$$

$$\hat{\theta}(\text{dollar}|\text{spam}) = \frac{\text{dollar}|\text{spam}}{\text{spam}} = \frac{1}{3}$$

6. Which of the following problems are more suited for a learning approach and which are more suited for a design approach? (from AML book)

Solution:

1. Determining the age at which a particular medical test should be performed

✓ A pattern exists ✓ Explanatory mathematical function does not exist ? There is data

This could be a good fit for a learning approach if enough data could be acquired. Patterns in effectiveness of things like medical tests are often difficult to tackle with design approaches so a learning approach could be useful. I don't know what kind of test it is or the parameters that they judge on, so I'm going to just assume that data is available and say this would be appropriate for a learning approach.

Either. Design approach will allow for experimental design and higher certainty but if data is available it may be faster and quicker to use a Learning approach.

2. Classifying numbers into primes and non-primes

✓ A pattern exists X Explanatory mathematical function does not exist ✓ There is data

This is clearly a design approach problem since primes, by their very nature, are mathematically defined.

Design: Math has ways of checking if a number is prime though finding prime numbers is slightly more challenging. Additionally, number theory suggests there might not be a pattern in terms of when prime numbers occur.

3. Detecting potential fraud in credit card charges

✓ A pattern exists ✓ Explanatory mathematical function does not exist ✓ There is data

This is an ideal candidate for a learning approach.

Learning: Data exists, there's no clear analytical solution

4. Determining the time it would take a falling object to hit the ground

✓ A pattern exists X Explanatory mathematical function does not exist ✓ There is data

Again, this is mathematically defined, so a design approach is appropriate.

Design: Physics has theories and laws that can easily and accurately predict this.

5. Determining the optimal cycle for traffic lights in a busy intersection

✓ A pattern exists ✓ Explanatory mathematical function does not exist ✓ There is data

This would be good for a learning approach since traffic activity is often difficult to mathematically define, but follows patterns.

Learning: data is cheap, system is not easily represented as an equation that has an optimal solution, there is likely a pattern in the data.

7. You have an unfair 6-sided die (i.e. 1 dice). Values 1 and 4 are rolled with probability 1/4. All other values are rolled with probability 1/8.

Solution:

(a) What is the formula for expected value in the discrete case?

$$E[X] = \sum_{i=1}^n x_i p(x_i)$$

(b) What is the expected value of a single throw of the biased die?

$$E[X] = \sum_{i=1}^6 Pr(d) * d = 1(\frac{1}{4}) + 2(\frac{1}{8}) + 3(\frac{1}{8}) + 4(\frac{1}{4}) + 5(\frac{1}{8}) + 6(\frac{1}{8}) = 3.25$$

(c) What is the variance of the biased die?

$$\sigma^2 = \sum_{i=1}^6 (x_i - \mu)^2 p_i$$

$$(1 - 3.25)^2(\frac{1}{4}) + (2 - 3.25)^2(\frac{1}{8}) + (3 - 3.25)^2(\frac{1}{8}) + (4 - 3.25)^2(\frac{1}{4}) + (5 - 3.25)^2(\frac{1}{8}) + (6 - 3.25)^2(\frac{1}{8})$$

$$\sigma^2 = \frac{47}{16} = 2.9375$$

8. Assume we know that $x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2)$ but we don't know the value of μ . Construct and solve for a likelihood function such that we maximize $P(x|\mu, \sigma)$ [Hint: you will need to know the analytical form of the normal equation, also assume x is iid. Hint 2: Computing the log likelihood is easiest.]

Solution:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n f(x_i, \mu, \sigma^2) \\ &= \frac{1}{\sqrt{\sigma^2 2\pi}^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \frac{1}{(\sigma^2 2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned} \tag{1}$$

$$\begin{aligned} \log L(\mu, \sigma^2) &= \log \frac{1}{(\sigma^2 2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= -\frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \end{aligned} \tag{2}$$

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2) = -\frac{1}{2\sigma^2} (2) \sum_{i=1}^n (x_i - \mu)(-1) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$\sum (x_i) - n\mu = 0 \implies \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

$$\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2}$$

$$-\frac{n}{2\sigma^2} + \frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2} = 0 \implies -n\sigma^2 + \sum (x_i - \bar{x})^2 = 0 \implies \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

TEACHER SOLUTION

There are two approaches Bayesian estimation and MLE.

The first solution is the MLE which assumes both μ and σ are unknown. Here we can write down the normal density function, convert it to log-likelihood, compute derivative and set it equal to 0. Solving that equation will tell us the MLE for μ and σ both.

We can also use Bayesian estimation where we assume μ is unknown but σ is known.

Solution 1:

$$\begin{aligned} L(\mu, \sigma) &= f(x|\mu, \sigma) = \prod_{i=1}^n (2\pi\sigma)^{-2} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ \ell(\mu, \sigma) &= \sum_{i=1}^n \left[-\log \sigma - \frac{1}{2} \log 2\pi - \frac{1}{2\sigma^2} (x_i - \mu)^2\right] \\ &= \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \right) \\ &= \sum_{i=1}^n \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \log \sqrt{2\sigma^2\pi} \right) \\ \frac{d \log \mathcal{L}(x|\mu, \sigma)}{d\mu} &= \sum_{i=1}^n \frac{-2 \cdot -1(x_i - \mu)}{2\sigma^2} = 0 \\ &= \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma^2} = \sum_{i=1}^n (x_i - \mu) \\ \sum_{i=1}^n x_i &= n \cdot \hat{\mu} \\ \frac{1}{n} \sum_{i=1}^n x_i &= \hat{\mu} \end{aligned}$$

Solution 2

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

also the Gaussian prior has a mean distribution that is itself normal: $p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(X)} = \alpha \times \prod_i^n p(x_i|\mu)p(\mu) \text{ and } \alpha = \frac{1}{p(X)}$$

Note: α is independent of μ .

$$\begin{aligned}
p(\mu|x) &= \alpha \times \prod_i^n \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right] \\
&\rightarrow \beta \times \exp\left[-1/2\left(\sum_i^n \left(\frac{(x_i - \mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\right)\right)\right] \\
&\rightarrow \gamma \times \exp\left[-1/2\left[(n/\sigma^2 + 1/\sigma_0^2)\mu^2 - 2(1/\sigma^2 \sum_i^n x_i - \mu_0/\sigma_0^2)\mu\right]\right] \\
&\rightarrow (n/\sigma^2 + 1/\sigma_0^2)\mu^2 - 2(1/\sigma^2 \sum_i^n x_i - \mu_0/\sigma_0^2)\mu + \gamma = \frac{(\mu - \mu_n)^2}{\sigma_n^2}
\end{aligned}$$

So from μ_n and σ_n^2 :

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \rightarrow \hat{\mu}_n = \frac{\sum_i^n x_i}{n} = \bar{X}$$

So our MLE for μ would (again) be $\frac{1}{n} \sum_{i=1}^n x_i$

9. Bias Variance Tradeoff

Solution: Squared bias is large in simple models, as flexibility increases the bias decreases as the model is more able to approximate the true underlying function. If bias is too high, the model is likely to underfit.

Variance increases as the model flexibility increases as the model is more able to account for weaker trends in the training data. One way to think about variance is a quantification of how the model would change if it was constructed using a different training data sample. If the model would change significantly, then the model is a high variance model and has likely overfit the training data.

Training error monotonically decreases as flexibility increases as more and more specific trends are being captured by the model. Testing data takes a u-shape where the left side is under-fitting, and the right side is over-fitting. Testing error is the sum of bias-squared plus variance.

10. From FML chapter 2. Let $X = R^2$. Consider the set of concepts of the form $c = (x, y) : x^2 + y^2 \leq r^2$ for some real number r . Show that this class can be (ϵ, δ) -PAC-learned from the training data of size $m \geq \frac{1}{\epsilon} \log \frac{1}{\delta}$.

Solution:

The generalization error is greater than ε with a probability less than δ

$$P[R(h) \leq \varepsilon] \leq 1 - \delta$$

$$e^{-\varepsilon m} \leq \delta$$

$$-\varepsilon m \leq \ln \delta$$

$$m \geq \frac{1}{\varepsilon} \ln \frac{1}{\delta}$$

The number of samples required for PAC learning:

$$m \geq \frac{1}{\varepsilon} (\ln |H| + \ln \frac{1}{\delta})$$

We know that since the concept class only contains a single hypothesis, we know that cardinality term drops out and leaves us with the form suggested in the question. $m \geq \frac{1}{\varepsilon} \ln \frac{1}{\delta}$

TEACHER SOLUTION

WLOG, suppose our target concept c is the circle around the origin with radius r . We will choose a slightly smaller radius s by

$$s := \inf\{s' : P(s' \leq \|x\| \leq r) < \varepsilon\}$$

Let A denote the annulus between radii s and r ; that is, $A := \{x : s \leq \|x\| \leq r\}$

By definition of s ,

$$P(A) \geq \varepsilon$$

In addition, our generalization error, $P(c\Delta L(T))$, must be small if T intersects A . We can state this as

$$P(c\Delta L(T)) > \varepsilon \implies T \cap A = \emptyset$$

Using $P(A) \geq \varepsilon$, we know that any point in T chosen according to p will “miss” region A with probability at most $1 - \varepsilon$. Defining error $:= P(c\Delta L(T))$, we can combine this with the above equation to see that

$$P(\text{error} > \varepsilon) \leq P(T \cap A = \emptyset) \leq (1 - \varepsilon)^m \leq e^{-m\varepsilon}$$

Setting δ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\varepsilon} \log(\frac{1}{\delta})$