# Bayes' rule and conditional probability

Ten problems on conditional probability and Bayes' rule: the product rule and the law of total probability, a diagnostic test at a 1% base rate, posterior odds as prior odds times a likelihood ratio, two tests in sequence, naive Bayes by hand, a counterexample separating independence from conditional independence, the Beta–Bernoulli posterior, precision as a posterior, and a calibrated score corrected for label shift, with worked solutions and the mistakes that swap a conditional or skip the normalizer.

**BEFORE YOU START**

Bayes' rule is one line of algebra, and most of the ways it goes wrong in practice come from a number that answers a different question: a test's sensitivity read as the chance that a positive result is right, a classifier's recall read as its precision, a score calibrated on balanced training data read as a probability in a population where positives are rare. These ten problems start from the definition of a conditional probability, build the product rule, the law of total probability and Bayes' rule from it, and then use them where an ML practitioner meets them: diagnostic tests, naive Bayes, a Beta prior on a click-through rate, precision and recall, and a classifier whose class balance changes at deployment. The five mistakes at the end are the ones that give a plausible number: a conditional read the wrong way round, a prior-times-likelihood score reported without normalizing, two correlated tests counted as independent evidence, independence taken to imply conditional independence, and a Beta posterior read off its exponents.

- $P(A)$ is the probability of the event $A$, $\neg A$ is its complement, with $P(\neg A) = 1 - P(A)$, and $A \cap B$ is the event that both happen.
- For $P(B) > 0$, the conditional probability of $A$ given $B$ is $P(A \mid B) = P(A \cap B)/P(B)$: the probability of $A$ once the outcomes outside $B$ have been ruled out and the rest rescaled to total 1. Given $B$, $P(\cdot \mid B)$ is itself a probability, so $P(\neg A \mid B) = 1 - P(A \mid B)$.
- Product rule: $P(A \cap B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$.
- Law of total probability: if $B_1, \ldots, B_m$ are disjoint and one of them always happens (a partition), then $P(A) = \sum_i P(A \mid B_i)\,P(B_i)$.
- Bayes' rule: $P(B \mid A) = P(A \mid B)\,P(B)/P(A)$, with $P(A)$ from the law of total probability. $P(B)$ is the prior, $P(A \mid B)$ the likelihood and $P(B \mid A)$ the posterior.
- The odds of $A$ are $O(A) = P(A)/P(\neg A)$, and $P(A) = O(A)/(1 + O(A))$.
- $A$ and $B$ are independent if $P(A \cap B) = P(A)\,P(B)$. They are conditionally independent given $C$ if $P(A \cap B \mid C) = P(A \mid C)\,P(B \mid C)$. For random variables, each holds for every combination of values.

- A diagnostic test for a condition $D$ has sensitivity $P(+ \mid D)$ and specificity $P(- \mid \neg D)$; the prevalence $P(D)$ is the base rate.
- The Beta distribution $\mathrm{Beta}(\alpha, \beta)$, with $\alpha, \beta > 0$, has density $\theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha,\beta)$ on $0 < \theta < 1$, where $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta$ makes it integrate to 1. Its mean is $\alpha/(\alpha+\beta)$.
- For a binary classifier with label $y$ and prediction $\hat{y}$: recall (the true positive rate) is $P(\hat{y}=1 \mid y=1)$, the false positive rate is $P(\hat{y}=1 \mid y=0)$, and precision is $P(y=1 \mid \hat{y}=1)$.

## Problems

**1.** In a mail corpus, $S$ is the event that a message is spam and $W$ the event that it contains the word *free*. The joint probabilities are $P(S \cap W) = 0.12$, $P(S \cap \neg W) = 0.18$, $P(\neg S \cap W) = 0.04$ and $P(\neg S \cap \neg W) = 0.66$. Compute $P(S)$, $P(W)$, $P(S \mid W)$ and $P(W \mid S)$, and check both forms of the product rule. difficulty ·

**2.** A model serves requests from three sources: web, with 50% of the traffic, mobile with 30% and an API with 20%. Its error rates on the three sources are 2%, 5% and 10%. What is its overall error rate, and what fraction of its errors come from the API? difficulty ·

**3.** A condition has prevalence 1%. A test for it has sensitivity 0.9 and specificity 0.95. A person tests positive. What is the probability that they have the condition? difficulty ··

**4.** Show that $O(D \mid +) = \dfrac{P(+ \mid D)}{P(+ \mid \neg D)}\, O(D)$: posterior odds are the likelihood ratio times the prior odds. Use it to recompute Problem 3, then find $P(D \mid -)$ for the same test. difficulty ··

**5.** After the positive result of Problem 3, a second test $T_2$ is run, with sensitivity 0.8 and specificity 0.9. The two tests are conditionally independent given the person's status: $P(T_1, T_2 \mid D) = P(T_1 \mid D)\,P(T_2 \mid D)$, and likewise given $\neg D$. Find $P(D \mid T_1+, T_2+)$ and $P(D \mid T_1+, T_2-)$, and show that updating on the two results one after the other gives the same answer as updating on both at once. difficulty ··

**6.** A naive Bayes spam filter uses three binary features, the presence of the words *free*, *invoice* and *meeting*, and the prior $P(\text{spam}) = 0.4$. Its estimated probabilities that each word $w$ is present are: difficulty ··

| **word** $w$ | $P(w \mid \text{spam})$ | $P(w \mid \text{ham})$ |
|---|---|---|
| *free* | 0.6 | 0.1 |
| *invoice* | 0.3 | 0.2 |
| *meeting* | 0.1 | 0.4 |

A message contains *free* and *meeting* but not *invoice*. Find $P(\text{spam} \mid x)$ for its feature vector $x$ and the class the filter assigns.

**7.** (a) $X$ and $Y$ are independent fair coin flips with values 0 and 1, and $Z = 1$ if $X = Y$, $Z = 0$ otherwise. Show that $X$ and $Y$ are independent but not conditionally independent given $Z$. (b) For the two tests of Problem 5, which are conditionally independent given $D$, compare $P(T_2{+} \mid T_1{+})$ with $P(T_2{+})$. difficulty ···

**8.** A click-through rate $\theta$ has prior $\text{Beta}(a, b)$. Of $N$ independent impressions, $k$ are clicked. Show that the posterior is a Beta distribution, give its parameters and its mean, and evaluate them for $a = b = 2$, $N = 10$, $k = 7$. difficulty ··

**9.** A fraud classifier has recall $r = 0.8$ and false positive rate $f = 0.05$. A fraction $\pi$ of transactions are fraudulent. Express the precision in terms of $r$, $f$ and $\pi$, and evaluate it at $\pi = 0.1$ and $\pi = 0.01$. difficulty ··

**10.** A classifier was trained where the positive rate is $\pi_{\text{tr}}$, and its score is calibrated there: $s(x) = P_{\text{tr}}(y = 1 \mid x)$. At deployment the positive rate is $\pi_{\text{te}}$, while the class-conditional distributions $p(x \mid y)$ are unchanged (label shift). Express $P_{\text{te}}(y = 1 \mid x)$ in terms of $s(x)$, $\pi_{\text{tr}}$ and $\pi_{\text{te}}$, and evaluate it for $\pi_{\text{tr}} = \tfrac{1}{2}$, $\pi_{\text{te}} = \tfrac{1}{10}$ and $s = 0.8$. difficulty ···

## Answers

1. $P(S) = 0.3$, $P(W) = 0.16$, $P(S \mid W) = \frac{3}{4}$, $P(W \mid S) = \frac{2}{5}$, and $P(W \mid S)P(S) = P(S \mid W)P(W) = 0.12 = P(S \cap W)$
2. $P(E) = 0.045$ and $P(\text{API} \mid E) = \frac{4}{9} \approx 0.444$
3. $P(D \mid +) = \frac{2}{13} \approx 0.154$
4. $O(D \mid +) = 18 \cdot \frac{1}{99} = \frac{2}{11}$, so $P(D \mid +) = \frac{2}{13}$; $O(D \mid -) = \frac{2}{19} \cdot \frac{1}{99} = \frac{2}{1881}$, so $P(D \mid -) = \frac{2}{1883} \approx 0.00106$
5. $P(D \mid T_1+, T_2+) = \frac{16}{27} \approx 0.593$ and $P(D \mid T_1+, T_2-) = \frac{4}{103} \approx 0.039$
6. $P(\text{spam} \mid x) = 0.0168/0.036 = \frac{7}{15} \approx 0.467$, so the filter assigns ham
7. $P(X = 1, Y = 1 \mid Z = 1) = \frac{1}{2} \neq \frac{1}{4} = P(X = 1 \mid Z = 1)P(Y = 1 \mid Z = 1)$, and $P(T_2+ \mid T_1+) = \frac{27}{130} \approx 0.208 \neq 0.107 = P(T_2+)$
8. The posterior is $\text{Beta}(a + k, b + N - k)$ with mean $\dfrac{a + k}{a + b + N}$; for $a = b = 2$, $N = 10$, $k = 7$ it is $\text{Beta}(9, 5)$ with mean $\frac{9}{14} \approx 0.643$
9. Precision $= \dfrac{r\pi}{r\pi + f(1 - \pi)}$: $0.64$ at $\pi = 0.1$ and $\frac{16}{115} \approx 0.139$ at $\pi = 0.01$
10. $P_{\text{te}}(y = 1 \mid x) = \dfrac{w_1 s}{w_1 s + w_0 (1 - s)}$ with $w_1 = \dfrac{\pi_{\text{te}}}{\pi_{\text{tr}}}$, $w_0 = \dfrac{1 - \pi_{\text{te}}}{1 - \pi_{\text{tr}}}$; for $\pi_{\text{tr}} = \frac{1}{2}$, $\pi_{\text{te}} = \frac{1}{10}$, $s = 0.8$ it is $\frac{4}{13} \approx 0.308$

# Worked solutions

## Problem 1

> In a mail corpus, $S$ is the event that a message is spam and $W$ the event that it contains the word *free*. The joint probabilities are $P(S \cap W) = 0.12$, $P(S \cap \neg W) = 0.18$, $P(\neg S \cap W) = 0.04$ and $P(\neg S \cap \neg W) = 0.66$. Compute $P(S)$, $P(W)$, $P(S \mid W)$ and $P(W \mid S)$, and check both forms of the product rule.

| | |
|---|---|
| $P(S) = 0.12 + 0.18 = 0.30$ and $P(W) = 0.12 + 0.04 = 0.16$. | $S$ is the disjoint union of $S \cap W$ and $S \cap \neg W$, and probabilities of disjoint events add; the same for $W$. |
| $P(S \mid W) = 0.12/0.16 = \frac{3}{4}$. | The definition: the joint probability divided by the probability of the event conditioned on. |
| $P(W \mid S) = 0.12/0.30 = \frac{2}{5}$. | The same numerator, now divided by $P(S)$, because the condition is $S$. |
| $P(W \mid S)\,P(S) = 0.4 \cdot 0.3 = 0.12$ and $P(S \mid W)\,P(W) = 0.75 \cdot 0.16 = 0.12$. | Each definition multiplied by its own denominator returns the joint probability $P(S \cap W)$. |
| $P(S) = 0.3$, $P(W) = 0.16$, $P(S \mid W) = \frac{3}{4}$, $P(W \mid S) = \frac{2}{5}$, **and** $P(W \mid S)P(S) = P(S \mid W)P(W) = 0.12 = P(S \cap W)$ | The two conditionals differ because they divide one joint probability by two different marginals (Mistake 1). Setting the two products of step 4 equal and dividing by $P(W)$ is Bayes' rule. |

## Problem 2

> A model serves requests from three sources: web, with 50% of the traffic, mobile with 30% and an API with 20%. Its error rates on the three sources are 2%, 5% and 10%. What is its overall error rate, and what fraction of its errors come from the API?

| | |
|---|---|
| Let $B_1, B_2, B_3$ be the events that a request comes from web, mobile or the API, and $E$ the event that the model errs. The $B_i$ form a partition, and the error rates are $P(E \mid B_i)$. | Every request comes from exactly one source, and an error rate on a source is a probability among that source's requests. |
| $P(E) = 0.5 \cdot 0.02 + 0.3 \cdot 0.05 + 0.2 \cdot 0.10 = 0.010 + 0.015 + 0.020 = 0.045.$ | Law of total probability: $E$ is the disjoint union of the $E \cap B_i$, and each has probability $P(E \mid B_i)P(B_i)$ by the product rule. |
| $P(B_3 \mid E) = \dfrac{P(E \mid B_3)\,P(B_3)}{P(E)} = \dfrac{0.020}{0.045} = \dfrac{4}{9}.$ | Bayes' rule; the numerator is the API's term of the sum in step 2. |
| $P(E) = 0.045$ **and** $P(\text{API} \mid E) = \frac{4}{9} \approx 0.444$ | The overall rate is a traffic-weighted average of the three rates, so it lies between 2% and 10%. The API carries 20% of the traffic and 44% of the errors: each source's share of the errors is its term of the sum divided by the total. |

## Problem 3

> A condition has prevalence 1%. A test for it has sensitivity 0.9 and specificity 0.95. A person tests positive. What is the probability that they have the condition?

| | |
|---|---|
| $P(D) = 0.01$, $P(+ \mid D) = 0.9$ and $P(+ \mid \neg D) = 1 - 0.95 = 0.05.$ | Given $\neg D$ the result is $+$ or $-$, so the false positive rate is one minus the specificity. |
| $P(+) = 0.9 \cdot 0.01 + 0.05 \cdot 0.99 = 0.009 + 0.0495 = 0.0585.$ | Law of total probability over the partition $D$, $\neg D$. |

| | |
|---|---|
| $P(D \mid +) = \dfrac{P(+ \mid D)\,P(D)}{P(+)} = \dfrac{0.009}{0.0585} = \dfrac{90}{585} = \dfrac{2}{13}.$ | Bayes' rule; $90 = 45 \cdot 2$ and $585 = 45 \cdot 13$. |
| $P(D \mid +) = \frac{2}{13} \approx 0.154$ | In 10,000 people, 100 have the condition and 90 of them test positive; of the 9,900 without it, 495 test positive. A positive result is a true one 90 times in 585. The false positives win because the group they come from is 99 times larger, which a sensitivity of 0.9 says nothing about (Mistake 1). |

## Problem 4

> Show that $O(D \mid +) = \dfrac{P(+ \mid D)}{P(+ \mid \neg D)}\,O(D)$: posterior odds are the likelihood ratio times the prior odds. Use it to recompute Problem 3, then find $P(D \mid -)$ for the same test.

| | |
|---|---|
| $P(D \mid +) = \dfrac{P(+ \mid D)P(D)}{P(+)}$ and $P(\neg D \mid +) = \dfrac{P(+ \mid \neg D)P(\neg D)}{P(+)}.$ | Bayes' rule for each of the two hypotheses, with the same evidence $+$. |
| $O(D \mid +) = \dfrac{P(D \mid +)}{P(\neg D \mid +)} = \dfrac{P(+ \mid D)}{P(+ \mid \neg D)} \cdot \dfrac{P(D)}{P(\neg D)}.$ | Divide the two lines of step 1: $P(+)$ is common to both and cancels, so the normalizer is never needed. |
| The likelihood ratio of a positive is $\mathrm{LR}_+ = 0.9/0.05 = 18$, and $O(D) = 0.01/0.99 = \frac{1}{99}$, so $O(D \mid +) = \frac{18}{99} = \frac{2}{11}$. | Step 2 with Problem 3's numbers. |
| $P(D \mid +) = \dfrac{2/11}{1 + 2/11} = \dfrac{2}{13}.$ | $P = O/(1 + O)$; it agrees with Problem 3. |

| | |
|---|---|
| $\mathrm{LR}_- = \dfrac{P(- \mid D)}{P(- \mid \neg D)} = \dfrac{0.1}{0.95} = \dfrac{2}{19}$, so $O(D \mid -) = \dfrac{2}{19} \cdot \dfrac{1}{99} = \dfrac{2}{1881}$ and $P(D \mid -) = \dfrac{2}{1883}$. | Step 2 holds for any evidence, here $-$; $P(- \mid D) = 1 - 0.9$ is the miss rate. Then $P = O/(1+O) = 2/(1881+2)$. |
| $O(D \mid +) = 18 \cdot \frac{1}{99} = \frac{2}{11}$, **so** $P(D \mid +) = \frac{2}{13}$; $O(D \mid -) = \frac{2}{19} \cdot \frac{1}{99} = \frac{2}{1881}$, **so** $P(D \mid -) = \frac{2}{1883} \approx 0.00106$ | The likelihood ratio is the test's whole contribution: 18 for a positive, $\frac{2}{19}$ for a negative. In log-odds, $\log O(D \mid +) = \log O(D) + \log \mathrm{LR}_+$, so evidence adds, which is the form a logistic regression's logit takes. |

## Problem 5

> After the positive result of Problem 3, a second test $T_2$ is run, with sensitivity 0.8 and specificity 0.9. The two tests are conditionally independent given the person's status: $P(T_1, T_2 \mid D) = P(T_1 \mid D)\,P(T_2 \mid D)$, and likewise given $\neg D$. Find $P(D \mid T_1+, T_2+)$ and $P(D \mid T_1+, T_2-)$, and show that updating on the two results one after the other gives the same answer as updating on both at once.

| | |
|---|---|
| $\dfrac{P(T_1+, T_2+ \mid D)}{P(T_1+, T_2+ \mid \neg D)} = \dfrac{0.9 \cdot 0.8}{0.05 \cdot 0.1} = 18 \cdot 8 = 144.$ | Conditional independence factorizes numerator and denominator, so the joint likelihood ratio is the product of $\mathrm{LR}_+ = 18$ for $T_1$ and $0.8/0.1 = 8$ for $T_2$. |
| $O(D \mid T_1+, T_2+) = 144 \cdot \frac{1}{99} = \frac{16}{11}$, so $P = \frac{16/11}{27/11} = \frac{16}{27}$. | Problem 4, step 2, with the pair of results as the evidence; then $P = O/(1+O)$. |
| One at a time: after $T_1+$, $O = \frac{2}{11}$ (Problem 4); using it as the prior for $T_2$ gives $\frac{2}{11} \cdot 8 = \frac{16}{11}$, the same. | The second update needs $P(T_2 \mid D, T_1+) = P(T_2 \mid D)$, which is what conditional independence states, so the posterior after $T_1$ can serve as the prior for $T_2$; the two factors then multiply in either order. |
| For $T_2-$: its likelihood ratio is $0.2/0.9 = \frac{2}{9}$, so $O = \frac{2}{11} \cdot \frac{2}{9} = \frac{4}{99}$ and $P = \frac{4}{103}$. | The miss rate is $1 - 0.8 = 0.2$ and the specificity is 0.9; then $P = 4/(99+4)$. |

| | |
|---|---|
| $P(D \mid T_1+, T_2+) = \frac{16}{27} \approx 0.593$ **and** $P(D \mid T_1+, T_2-) = \frac{4}{103} \approx 0.039$ | Two positives turn 1% into better than even, and a negative second test undoes most of the first positive. Without conditional independence the likelihood ratios cannot be multiplied (Mistake 3). |

## Problem 6

A naive Bayes spam filter uses three binary features, the presence of the words *free*, *invoice* and *meeting*, and the prior $P(\text{spam}) = 0.4$. Its estimated probabilities that each word $w$ is present are:

| **word** $w$ | $P(w \mid \text{spam})$ | $P(w \mid \text{ham})$ |
|---|---|---|
| *free* | 0.6 | 0.1 |
| *invoice* | 0.3 | 0.2 |
| *meeting* | 0.1 | 0.4 |

A message contains *free* and *meeting* but not *invoice*. Find $P(\text{spam} \mid x)$ for its feature vector $x$ and the class the filter assigns.

| | |
|---|---|
| $P(x \mid c) = \prod_j P(x_j \mid c)$ for each class $c$. | The naive Bayes assumption: the features are conditionally independent given the class. |
| $P(x \mid \text{spam}) = 0.6 \cdot (1 - 0.3) \cdot 0.1 = 0.042$ and $P(x \mid \text{ham}) = 0.1 \cdot (1 - 0.2) \cdot 0.4 = 0.032$. | An absent word is evidence too: its factor is the probability of absence, one minus the table entry. |
| $P(\text{spam}, x) = 0.4 \cdot 0.042 = 0.0168$ and $P(\text{ham}, x) = 0.6 \cdot 0.032 = 0.0192$. | Product rule, prior times likelihood, with $P(\text{ham}) = 0.6$. |
| $P(x) = 0.0168 + 0.0192 = 0.036$. | Law of total probability over the two classes. |

| | |
|---|---|
| $P(\text{spam} \mid x) = 0.0168/0.036 = \frac{7}{15} \approx 0.467,$ **so the filter assigns ham** | Bayes' rule: step 3 divided by step 4. In odds, the likelihood ratio $0.042/0.032 = \frac{21}{16}$ favors spam, but the prior odds $\frac{2}{3}$ favor ham, and the product $\frac{7}{8}$ is below 1. The step 3 numbers are not probabilities of anything given $x$ until they are divided by their sum (Mistake 2). |

## Problem 7

> (a) $X$ and $Y$ are independent fair coin flips with values 0 and 1, and $Z = 1$ if $X = Y$, $Z = 0$ otherwise. Show that $X$ and $Y$ are independent but not conditionally independent given $Z$. (b) For the two tests of Problem 5, which are conditionally independent given $D$, compare $P(T_2{+} \mid T_1{+})$ with $P(T_2{+})$.

| | |
|---|---|
| $P(X = x, Y = y) = \frac{1}{4} = P(X = x)\,P(Y = y)$ for all four pairs $(x, y)$. | Fair independent flips make the four outcomes equally likely, and each marginal is $\frac{1}{2}$. |
| $P(Z = 1) = P(X = Y) = \frac{1}{2}.$ | $Z = 1$ on two of the four outcomes, $(0,0)$ and $(1,1)$. |
| $P(X = 1, Y = 1 \mid Z = 1) = \dfrac{1/4}{1/2} = \dfrac{1}{2}.$ | The outcome $(1,1)$ lies inside $\{Z = 1\}$, so its intersection with that event is itself. |
| $P(X = 1 \mid Z = 1) = P(Y = 1 \mid Z = 1) = \dfrac{1/4}{1/2} = \dfrac{1}{2}$, whose product is $\frac{1}{4}$. | Within $\{Z = 1\}$, $X = 1$ happens only at $(1,1)$, and the same for $Y$. |
| $P(T_2{+}) = 0.8 \cdot 0.01 + 0.1 \cdot 0.99 = 0.107.$ | Law of total probability over $D$, $\neg D$, with $P(T_2{+} \mid \neg D) = 1 - 0.9$. |
| $P(T_1{+}, T_2{+}) = 0.9 \cdot 0.8 \cdot 0.01 + 0.05 \cdot 0.1 \cdot 0.99 = 0.0072 + 0.00495 = 0.01215,$ so $P(T_2{+} \mid T_1{+}) = 0.01215/0.0585 = \frac{27}{130}.$ | Total probability again, with the conditional independence of Problem 5 factorizing each branch; $P(T_1{+}) = 0.0585$ is Problem 3, step 2. |

| | |
|---|---|
| $P(X=1, Y=1 \mid Z=1) = \frac{1}{2} \neq \frac{1}{4} = P(X=1 \mid Z=1)P(Y=1 \mid Z=1)$, **and** $P(T_2{+} \mid T_1{+}) = \frac{27}{130} \approx 0.208 \neq 0.107 = P(T_2{+})$ | Neither property implies the other. In (a), conditioning on a common effect couples two independent causes: once $X = Y$ is known, $X$ determines $Y$, and this is why independent features do not stay independent given a label (Mistake 4). In (b), a common cause couples its effects: a first positive raises the chance of $D$, which raises the chance of a second positive. |

## Problem 8

> A click-through rate $\theta$ has prior $\mathrm{Beta}(a, b)$. Of $N$ independent impressions, $k$ are clicked. Show that the posterior is a Beta distribution, give its parameters and its mean, and evaluate them for $a = b = 2$, $N = 10$, $k = 7$.

| | |
|---|---|
| $p(\theta \mid \text{data}) \propto P(\text{data} \mid \theta)\, p(\theta) = \theta^{k}(1-\theta)^{N-k} \cdot \dfrac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}.$ | Bayes' rule for a continuous parameter: the denominator $p(\text{data})$ does not depend on $\theta$. Each click contributes a factor $\theta$ and each non-click $1-\theta$; the order of the outcomes is fixed, so no binomial coefficient is needed, and it would be a constant anyway. |
| $p(\theta \mid \text{data}) \propto \theta^{a+k-1}(1-\theta)^{b+N-k-1}.$ | Powers of the same base add their exponents; $B(a,b)$ is a constant. |
| The posterior is $\mathrm{Beta}(a+k,\, b+N-k)$. | A $\mathrm{Beta}(\alpha,\beta)$ density is $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ times a constant, so the exponents in step 2 are the parameters minus one; the constant is whatever makes the density integrate to 1, so it must be $1/B(a+k, b+N-k)$. |
| Its mean is $\dfrac{a+k}{(a+k)+(b+N-k)} = \dfrac{a+k}{a+b+N}.$ | The Beta mean $\alpha/(\alpha+\beta)$; the $k$ cancels in the denominator. |

| | |
|---|---|
| With $a = b = 2$, $N = 10$, $k = 7$: $\mathrm{Beta}(9, 5)$, with mean $\frac{9}{14}$. | $2 + 7 = 9$, $2 + 10 - 7 = 5$, and $9/(9 + 5)$. |
| **The posterior is** $\mathrm{Beta}(a + k, b + N - k)$ **with mean** $\dfrac{a + k}{a + b + N}$**; for** $a = b = 2$, $N = 10$, $k = 7$ **it is** $\mathrm{Beta}(9, 5)$ **with mean** $\frac{9}{14} \approx 0.643$ | The mean is $\frac{a+b}{a+b+N} \cdot \frac{a}{a+b} + \frac{N}{a+b+N} \cdot \frac{k}{N}$, a weighted average of the prior mean $\frac{1}{2}$ and the maximum likelihood estimate $\frac{7}{10}$, so the prior acts like $a + b$ extra impressions. As $N$ grows the data's weight goes to 1. |

## Problem 9

> A fraud classifier has recall $r = 0.8$ and false positive rate $f = 0.05$. A fraction $\pi$ of transactions are fraudulent. Express the precision in terms of $r$, $f$ and $\pi$, and evaluate it at $\pi = 0.1$ and $\pi = 0.01$.

| | |
|---|---|
| $P(\hat{y} = 1 \mid y = 1) = r$, $P(\hat{y} = 1 \mid y = 0) = f$ and $P(y = 1) = \pi$. | Recall and the false positive rate condition on the true label: they are likelihoods. Precision conditions on the prediction: it is a posterior. |
| $P(\hat{y} = 1) = r\pi + f(1 - \pi)$. | Law of total probability over $y = 1$ and $y = 0$. |
| $P(y = 1 \mid \hat{y} = 1) = \dfrac{r\pi}{r\pi + f(1 - \pi)}$. | Bayes' rule with step 2 as the denominator. |
| At $\pi = 0.1$: $\dfrac{0.08}{0.08 + 0.045} = \dfrac{0.08}{0.125} = 0.64$. At $\pi = 0.01$: $\dfrac{0.008}{0.008 + 0.0495} = \dfrac{0.008}{0.0575} = \dfrac{16}{115}$. | Step 3 with the numbers; $8/57.5 = 16/115$. |

| | |
|---|---|
| **Precision** $= \dfrac{r\pi}{r\pi + f(1-\pi)}$: $0.64$ **at** $\pi = 0.1$ **and** $\frac{16}{115} \approx 0.139$ **at** $\pi = 0.01$ | Recall and false positive rate are properties of the classifier on each class and do not move with $\pi$; precision does. This is Problem 3 again: precision is the positive predictive value, and at a low base rate the false positives from the large negative class dominate. A model evaluated on a balanced test set reports a precision it will not have in production. |

## Problem 10

> A classifier was trained where the positive rate is $\pi_{\text{tr}}$, and its score is calibrated there: $s(x) = P_{\text{tr}}(y = 1 \mid x)$. At deployment the positive rate is $\pi_{\text{te}}$, while the class-conditional distributions $p(x \mid y)$ are unchanged (label shift). Express $P_{\text{te}}(y = 1 \mid x)$ in terms of $s(x)$, $\pi_{\text{tr}}$ and $\pi_{\text{te}}$, and evaluate it for $\pi_{\text{tr}} = \frac{1}{2}$, $\pi_{\text{te}} = \frac{1}{10}$ and $s = 0.8$.

| | |
|---|---|
| $\dfrac{s}{1-s} = \dfrac{p(x \mid 1)}{p(x \mid 0)} \cdot \dfrac{\pi_{\text{tr}}}{1-\pi_{\text{tr}}}.$ | Problem 4, step 2, under the training distribution, with the observation $x$ as the evidence and densities in place of probabilities. |
| $O_{\text{te}}(x) = \dfrac{P_{\text{te}}(y=1 \mid x)}{P_{\text{te}}(y=0 \mid x)} = \dfrac{p(x \mid 1)}{p(x \mid 0)} \cdot \dfrac{\pi_{\text{te}}}{1-\pi_{\text{te}}}.$ | The same identity at deployment; label shift means the likelihood ratio $p(x \mid 1)/p(x \mid 0)$ is the one in step 1. |
| $O_{\text{te}}(x) = \dfrac{w_1 s}{w_0(1-s)}$, with $w_1 = \pi_{\text{te}}/\pi_{\text{tr}}$ and $w_0 = (1-\pi_{\text{te}})/(1-\pi_{\text{tr}})$. | Divide step 2 by step 1: the unknown likelihood ratio cancels, and the classifier's score is the only access to it that is needed. |
| $P_{\text{te}}(y = 1 \mid x) = \dfrac{w_1 s}{w_1 s + w_0(1-s)}.$ | $P = O/(1+O)$, with numerator and denominator multiplied by $w_0(1-s)$. |
| $w_1 = \frac{1}{5}$, $w_0 = \frac{9}{5}$, so $P_{\text{te}} = \dfrac{0.16}{0.16 + 0.36} = \dfrac{4}{13}.$ | $\frac{1}{5} \cdot 0.8 = 0.16$ and $\frac{9}{5} \cdot 0.2 = 0.36$. |

| | |
|---|---|
| $P_{\text{te}}(y=1 \mid x) = \dfrac{w_1 s}{w_1 s + w_0(1-s)}$ **with** $w_1 = \dfrac{\pi_{\text{te}}}{\pi_{\text{tr}}}$, $w_0 = \dfrac{1-\pi_{\text{te}}}{1-\pi_{\text{tr}}}$; **for** $\pi_{\text{tr}} = \frac{1}{2}$, $\pi_{\text{te}} = \frac{1}{10}$, $s = 0.8$ **it is** $\frac{4}{13} \approx 0.308$ | Each class's score is reweighted by its prior ratio and the two are renormalized. At these priors $O_{\text{te}} = O_{\text{tr}}/9$, so $P_{\text{te}} > \frac{1}{2}$ needs $s > 0.9$: the 0.5 threshold chosen on balanced data flags many more positives than the deployed posterior supports. The correction is only as good as the calibration of $s$ and the label-shift assumption. |

## Where this goes wrong

### 1. Sensitivity read as the chance that a positive is right

A test described as “90% accurate” invites reading the 90% as the chance that a result is right, and the conditional bar does not show which way round it points.

| | |
|---|---|
| $P(+ \mid D) = 0.9$, $P(+ \mid \neg D) = 0.05$ and $P(D) = 0.01$ | Right so far: the set-up of Problem 3. |
| “The test catches 90% of cases, so a positive means a 90% chance of having the condition.” | The shortcut that causes the mistake: it swaps $P(+ \mid D)$ for $P(D \mid +)$ and so never uses the base rate. |
| $P(D \mid +) = 0.9$ ✗ | $P(D \mid +) = \frac{2}{13} \approx 0.154$ (Problem 3): the two conditionals are equal only when $P(D) = P(+)$, and here $P(+) = 0.0585$ is almost six times $P(D)$. The ML version is reading recall as precision (Problem 9). |

### 2. Naive Bayes scores reported without normalizing

Bayes' rule is often remembered as “posterior is prior times likelihood”, and the proportionality sign is easy to drop.

| | |
|---|---|
| $P(\text{spam})\,P(x \mid \text{spam}) = 0.4 \cdot 0.042 = 0.0168$ | Right so far: Problem 6, step 3. |

| | |
|---|---|
| “The posterior is the prior times the likelihood.” | The shortcut that causes the mistake: that is the posterior only up to the factor $1/P(x)$, which is the same for every class and so is easy to forget. |
| $P(\text{spam} \mid x) = 0.0168$ ✗ | That is the joint $P(\text{spam}, x)$. Dividing by $P(x) = 0.036$ gives $\frac{7}{15}$ (Problem 6). Comparing unnormalized scores still picks the right class, which hides the slip, but any threshold other than the argmax, or any calibration plot, needs the normalized value. |

### 3. Repeating the same test counted as independent evidence

Problem 5 multiplied two likelihood ratios, and repeating a positive test looks like a cheap way to get a second one.

| | |
|---|---|
| $O(D \mid T_1{+}) = 18 \cdot \frac{1}{99} = \frac{2}{11}$ | Right so far: Problem 4, step 3. |
| “Run the same test again on the same sample; a second positive multiplies the odds by 18 again.” | The shortcut that causes the mistake: Problem 5's multiplication used conditional independence given $D$, and a repeat on the same sample tends to reproduce whatever caused the first result, false positives included. |
| $O(D \mid T_1{+}, T_1'{+}) = 18 \cdot 18 \cdot \frac{1}{99} = \frac{36}{11}$, so $P = \frac{36}{47} \approx 0.766$ ✗ | If a false positive comes from something in the sample, the repeat reproduces it with probability near 1, its likelihood ratio given the first result is near 1, and the posterior stays near $\frac{2}{13}$. The same error inflates a naive Bayes model's confidence when two features are near-copies of each other. |

## 4. Conditional independence inferred from independence

Independence is the property usually checked in the data, and naive Bayes needs conditional independence, so it is tempting to treat one as evidence for the other.

| | |
|---|---|
| $P(X = x, Y = y) = P(X = x)\,P(Y = y)$ for all $x, y$ | Right so far: Problem 7, step 1. |
| "$X$ and $Y$ are independent, so knowing $Z$ cannot make them dependent." | The shortcut that causes the mistake: conditioning on a common effect of $X$ and $Y$ couples them, even when nothing else does. |
| $P(X = 1, Y = 1 \mid Z = 1) = P(X = 1 \mid Z = 1)\,P(Y = 1 \mid Z = 1) = \frac{1}{4}$ ✗ | The left side is $\frac{1}{2}$ (Problem 7). A naive Bayes model on two independent features with the label $Z$ makes exactly this assumption and predicts $P(Z = 1 \mid x) = \frac{1}{2}$ for every input, although each input determines $Z$. |

## 5. Beta posterior parameters read off the exponents, one too small

The posterior density is written as a product of powers, and the exponents look like the parameters.

| | |
|---|---|
| $p(\theta \mid \text{data}) \propto \theta^{a+k-1}(1-\theta)^{b+N-k-1}$ | Right so far: Problem 8, step 2. |
| "The posterior is a Beta whose parameters are the two exponents." | The shortcut that causes the mistake: a $\text{Beta}(\alpha, \beta)$ density has exponents $\alpha - 1$ and $\beta - 1$, so each exponent is one less than its parameter. |
| The posterior is $\text{Beta}(a + k - 1, b + N - k - 1) = \text{Beta}(8, 4)$, with mean $\frac{8}{12} = \frac{2}{3}$ ✗ | The posterior is $\text{Beta}(9, 5)$ with mean $\frac{9}{14}$ (Problem 8). The wrong mean $\frac{a+k-1}{a+b+N-2}$ is the posterior mode, the MAP estimate, so the slip passes for a sensible number. With the uniform prior $a = b = 1$ it gives $\text{Beta}(k, N - k)$, which is not a distribution at all when $k = 0$. |

**More at eulerpad.app.** This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

**EulerPad**, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 2 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/bayes-rule-and-conditional-probability