Alpaydin's Book: Ex. 14.6.2 BL

Machine Learning 2024-25 Course Activity

Furno Francesco - francesco.furno@studenti.unipd.it - 2139507

December 2, 2024

Let us denote by x the number of spam emails I receive in a random sample of n. Assume that the prior for q, the proportion of spam emails is uniform in $[0, 1]$. Find the posterior distribution for $p(q|x)$.

Overview of the problem

This problem describes a situation where we observe some independent events, each of which can have only two possible outcomes, Y/N. We can represent each event with independent random variables where:

- $X = 1$, if X is a spam email
- $X = 0$ otherwise

The problem can be easily represented with a binomial distribution. In fact, the binomial distribution describes the probability to observe x successes over n independent events with success rate of q. The binomial formula is the following:

$$P(X = x \mid q) = \binom{n}{x} \cdot q^x (1-q)^{n-x}$$

where:

- $P(X = x \mid q)$ is the probability of observing x spam emails given q;
- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient, which counts the number of way it is possible to find x spam emails over n;
- q^x is the probability of x successes, where an email is spam;
- $(1-q)^{n-x}$ is the probability of $n - x$ failures, where the emails are not spam.

Bayes Formula

The posterior distribution for q given x can be calculated using the Bayes theorem:

$$P(q \mid x) = \frac{P(x \mid q) \cdot P(q)}{P(x)}$$

where:

- $P(q)$ is the prior, which expresses our initial informations about the training data.

Since $P(x)$ does not depend on q, we can consider it as a constant which we can remove from the equation. It follows that:

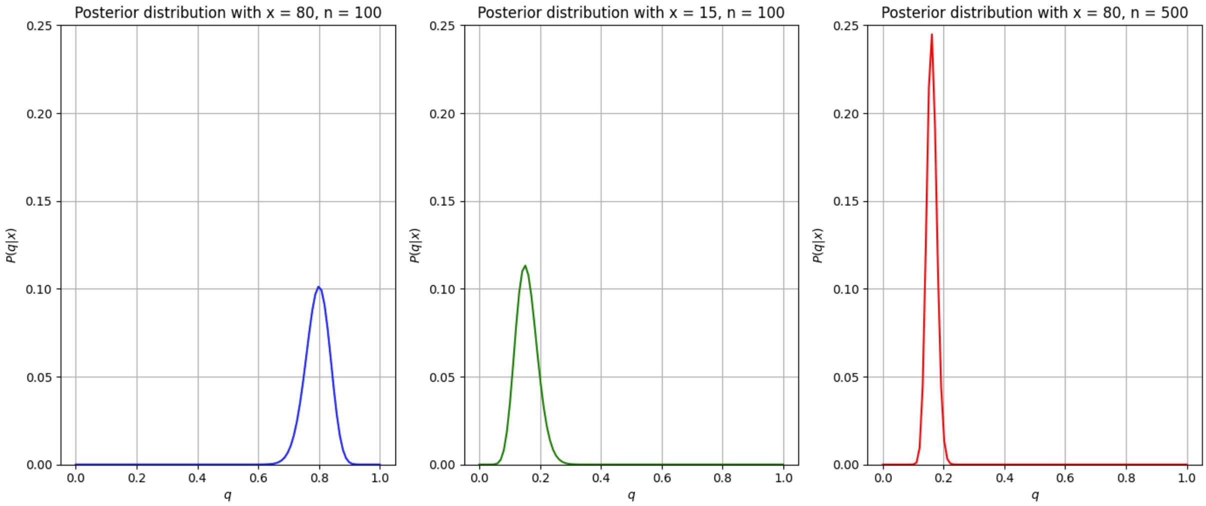
$$P(q \mid x) \propto P(x \mid q) \cdot P(q)$$

In this case we know that the proportion of spam emails $P(q)$ is uniform in $[0, 1]$: the proportional distribution will be normalized to ensure it integrates to 1 in the range $[0, 1]$.

Plots

Let's see how the posterior distribution changes with different parameters, in particular, what happens if:

- we increase the number of observed spam emails x ;
- we increase the total number of emails n



When x (the number of spam emails) is high, the posterior distribution becomes concentrated around high values of q, the probability that an email is spam.

When x is low, the posterior distribution is concentrated around low values of q, the probability that an email is not spam.

Additionally, as n (the total number of emails) increases, the posterior distribution becomes more peaked, reflecting reduced uncertainty about the true value of q.