Quantifying Uncertainty of Sensitivity and Specificity Metrics

Noah Ripstein

May 29, 2023

Sensitivity and specificity reported using Scikit-Learn's built in metrics are useful, but do present all of the relevant information regarding sensitivity and specificity of a classification model. In particular, those single number metrics do not provide an account of the uncertainty associated with reporting a single number. This is where further statistical analysis can be used to quantify the uncertainty in the prediction.

Quantifying uncertainty is especially valuable in cases where the testing data sample size is relatively small, and it becomes even more critical when there is a scarcity of data. By assessing the uncertainty, we gain insight into the range of plausible parameter values supported by the limited data, allowing for a more comprehensive interpretation of the results.

We can view the problem of determining sensitivity and specificity as a task of parameter estimation. A Bayesian approach lends itself to a straightforward method for this task. Below are the mathematical steps required to generate a PDF for sensitivity. The procedure for specificity is similar.

There are 4 components of a Bayesian parameter estimation problem which must be clearly stated:

1. The hypotheses (must be mutually exclusive and exhaustive) $H_1, H_2, ..., H_n$.
2. The prior probabilities of each hypothesis $P(H)$.
3. The data.

4. The likelihood function $P(D|H)$.

Hypotheses:

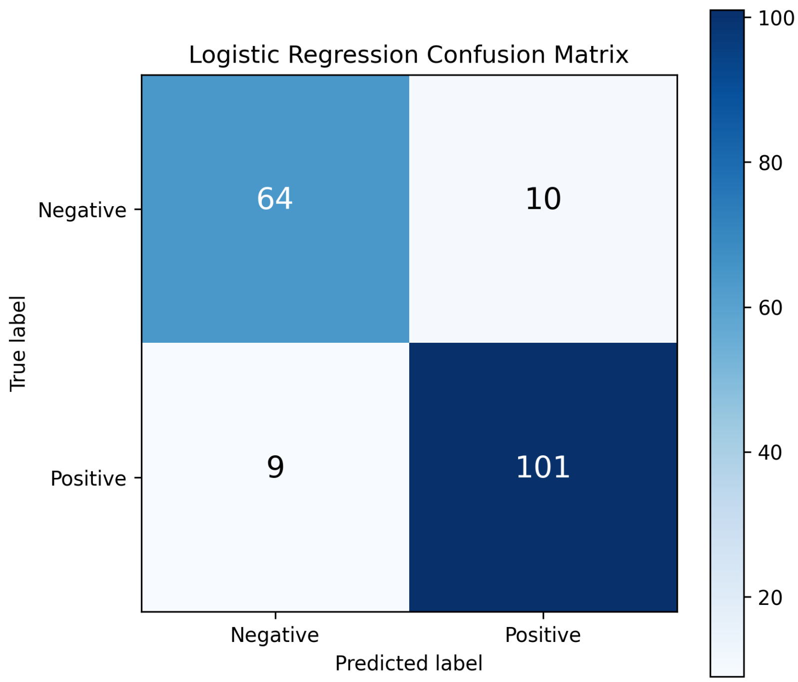
Let's call sensitivity, the parameter we're trying to estimate, ψ. We know that $\psi \in [0, 1]$. We can therefore create 100 independent, discrete hypotheses. This is achieved using binning. For simplicity, I will treat Hypotheses as holding that $\psi =$ the midpoint of the bin's range.
Hypothesis 1: $\psi \in [0, 0.01)$
Hypothesis 2: $\psi \in [0.01, 0.02)$
Hypothesis 3: $\psi \in [0.02, 0.03)$
...
Hypothesis 99: $\psi \in [0.98, 0.99)$
Hypothesis 100: $\psi \in [0.99, 1]$

Priors:

Before seeing the model's results, we don't have any *apriori* belief for what the value of ψ is, so we can assign uniform priors:
$P(\psi) = \frac{1}{100} \forall \psi$

Data:

The data relevant to assessing sensitivity and specificity comes from a confusion matrix. Sensitivity is defined to be the probability of a the model predicting someone has heart disease when they truly have it, and specificity is the probability that the model predicts someone doesn't have heart disease given that they do not have it. The confusion matrix below, which is from my baseline logistic regression model, serves as an example:



The data relevant for calculating sensitivity, ψ, is the bottom row of the confusion matrix: There are $n = 110$ samples with heart disease, $k = 101$ of which were correctly predicted. We will label this data D.

Likelihood Function:

We can view n classifications (which can only be positive or negative) as Bernoulli events with k positive predictions, so the probability of observing the data D given a hypothesized value of ψ is given by:

$$P(D|\psi) = \binom{n}{k}(\psi)^n(1-\psi)^{n-k}$$

$$P(D|\psi) = \binom{110}{101}(\psi)^{101}(1-\psi)^{9}$$

Posterior Probability:

From the definition of Bayes' Theorem, we have:

$$P(\psi|D) = \frac{P(D|\psi)P(\psi)}{\sum_{i=1}^{100} P(D|\psi_i)P(\psi_i)}$$

Which can be simplified because the uniform priors

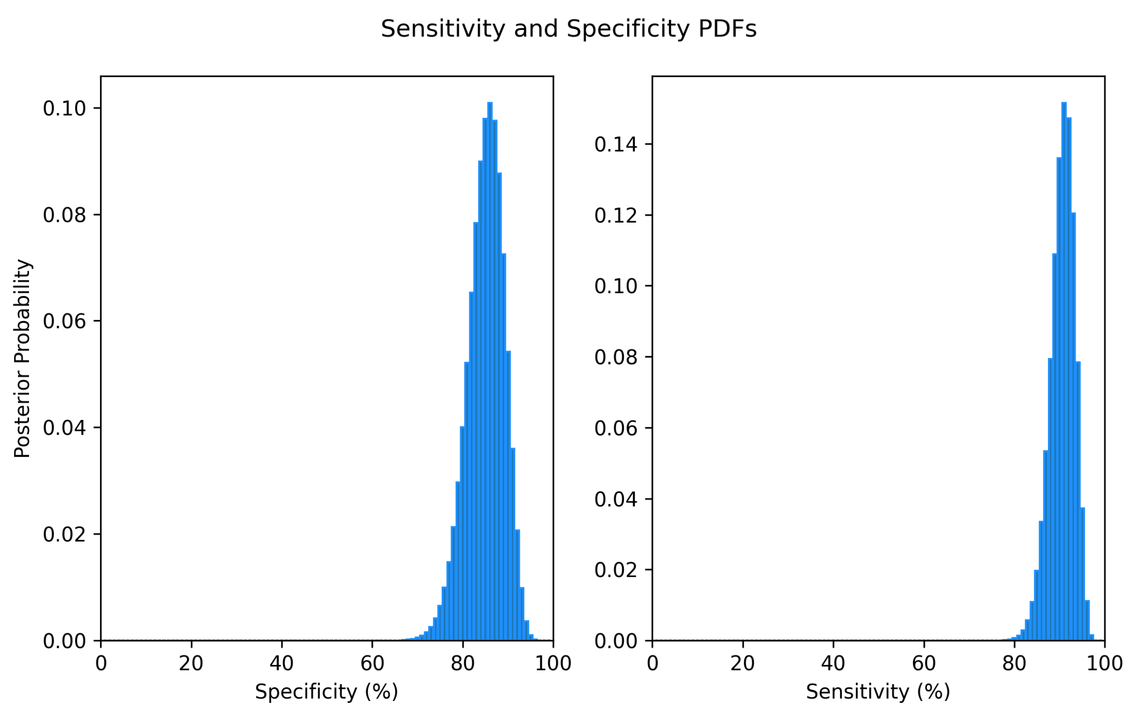
$$P(\psi|D) = \frac{P(D|\psi)}{\sum_{i=1}^{100} P(D|\psi_i)}$$

$$P(\psi|D) = \frac{\binom{110}{101}(\psi)^{101}(1-\psi)^9}{\sum_{i=1}^{100}\binom{110}{101}(\psi_i)^{101}(1-\psi_i)^9}$$

Which can be simplified because $\binom{110}{101}$ appears in the numerator and denominator

$$P(\psi|D) = \frac{(\psi)^{101}(1-\psi)^9}{\sum_{i=1}^{100}(\psi_i)^{101}(1-\psi_i)^9}$$

After evaluating the posterior probability for each value of ψ, the process can be repeated for specificity (in which case the data would be taken from the top row of the confusion matrix, so we would have $n = 74$, $k = 64$). We can plot the posterior probabilities to construct a probability distribution function (PDF), which will enable us to visualize the uncertainty.



In this example, the PDFs have the following characteristics:

Metric	**Mode**	**95% Credible Interval**
Sensitivity	86.5	(78.5, 92.5)
Specificity	91.5	(86.5, 95.5)

Algorithm used to Find the 95% Credible Interval From the Posterior Distribution

1. **Ordering the Bins**: The bins representing the parameter values were sorted in descending order based on their posterior probabilities. This ordering facilitates the identification of the highest posterior probability values.

2. **Accumulating Posteriors**: Starting from the bin with the highest posterior probability, the posterior probabilities were cumulatively summed until the accumulated sum reached 95% of the total posterior probability.

3. **Determining the Interval Bounds**: The lowest and highest parameter values associated with the summed posterior probabilities were identified as the bounds of the 95% credible interval. These bounds represent the range of parameter values within which 95% of the posterior probability is concentrated.

Please note that this algorithm assumes that the posterior distribution is unimodal. (This will be true for sensitivity and specificity parameter estimation problems like this one). If the posterior distribution is multimodal, then alternative methods may be necessary to estimate the credible interval effectively.