Bayesian Classification

Advanced Statistical Inference

Simone Rossi

Logistic Regression

1. Explain why a linear model $\boldsymbol{w}^\top \boldsymbol{x}$ is not sufficient by itself for binary classification if we want probabilistic predictions. Why does the sigmoid function solve this issue?

2. For logistic regression,

$$P(y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x})}.$$

   Given

$$\boldsymbol{w} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

   compute $\boldsymbol{w}^\top \boldsymbol{x}$, $P(y = 1 \mid \boldsymbol{x}, \boldsymbol{w})$, and $P(y = 0 \mid \boldsymbol{x}, \boldsymbol{w})$.

3. Show that the Bernoulli likelihood for one labeled example can be written as

$$p(y_n \mid \boldsymbol{w}, \boldsymbol{x}_n) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)^{y_n} \left(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)\right)^{1-y_n}.$$

4. Consider a dataset with two observations:

$$(\boldsymbol{x}_1, y_1) = ((1, 0)^\top, 1), \qquad (\boldsymbol{x}_2, y_2) = ((1, 1)^\top, 0),$$

   and parameter vector

$$\boldsymbol{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

   Compute:

   - $\sigma(\boldsymbol{w}^\top \boldsymbol{x}_1)$ and $\sigma(\boldsymbol{w}^\top \boldsymbol{x}_2)$;
   - the likelihood of each data point;
   - the joint log-likelihood of the dataset.

Bayesian Logistic Regression

1. Assume the prior
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \sigma_w^2 \boldsymbol{I}).$$
Write the posterior distribution
$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{X})$$
up to proportionality. Why is this posterior not available in closed form?

2. Show that maximum a posteriori estimation for Bayesian logistic regression is equivalent to minimizing
$$\mathcal{U}(\boldsymbol{w}) = -\log p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}) - \log p(\boldsymbol{w}).$$
What role does the prior play in this optimization problem?

3. Consider a one-dimensional logistic regression model with scalar parameter w, one observation $(x, y) = (2, 1)$, and prior
$$p(w) = \mathcal{N}(w \mid 0, 1).$$
Compute the log posterior up to an additive constant:
$$\log p(w \mid y, x) = \log p(y \mid w, x) + \log p(w) + \text{const}.$$
Evaluate this expression at $w = 0$ and $w = 1$ and say which value is larger.

4. Give one reason why each of the following can be used for Bayesian logistic regression, and one tradeoff:
   - MAP estimation;
   - Laplace approximation;
   - variational inference;
   - MM.

Prediction and Uncertainty

1. Suppose a Gaussian approximation to the posterior is
$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\mu} = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.1 \end{pmatrix}.$$
For
$$\boldsymbol{x}_\star = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$
compute:

- the mean $\boldsymbol{\mu}^\top \boldsymbol{x}_\star$ of the latent score $f_\star$;
- the variance $\boldsymbol{x}_\star^\top \boldsymbol{\Sigma} \boldsymbol{x}_\star$;
- the approximate predictive probability using

$$\sigma\left(\frac{\boldsymbol{\mu}^\top \boldsymbol{x}_\star}{\sqrt{1 + \frac{\pi}{8} \boldsymbol{x}_\star^\top \boldsymbol{\Sigma} \boldsymbol{x}_\star}}\right).$$

2. In Bayesian classification, what is the difference between:
   - a point prediction for the class label;
   - a predictive probability;
   - epistemic uncertainty;
   - aleatoric uncertainty?
3. Suppose posterior samples give the following predictive probabilities for the positive class on a test point:

$$0.70,\ 0.80,\ 0.65,\ 0.75.$$

   Use Monte Carlo integration to estimate $P(y_\star = 1 \mid \boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{X})$. What does the spread of these four values suggest qualitatively about uncertainty?

Evaluation

1. A classifier on a binary test set yields the confusion matrix counts

$$\text{TP} = 18, \qquad \text{TN} = 70, \qquad \text{FP} = 6, \qquad \text{FN} = 6.$$

   Compute:
   - accuracy;
   - precision;
   - recall;
   - F1 score.
2. Explain why accuracy can be misleading on imbalanced datasets. Give a concrete example of a classifier with high accuracy but poor practical usefulness.
3. Suppose we have three posterior samples for the probability of the true test label:

$$p(y_\star \mid \boldsymbol{w}^{(1)}, \boldsymbol{x}_\star) = 0.9, \qquad p(y_\star \mid \boldsymbol{w}^{(2)}, \boldsymbol{x}_\star) = 0.6, \qquad p(y_\star \mid \boldsymbol{w}^{(3)}, \boldsymbol{x}_\star) = 0.3.$$

   Compute the Monte Carlo approximation of the predictive test log-likelihood

$$\log p(y_\star \mid \boldsymbol{x}_\star, \boldsymbol{y}, \boldsymbol{X}) \approx \log\left(\frac{1}{3}\sum_{s=1}^{3} p(y_\star \mid \boldsymbol{w}^{(s)}, \boldsymbol{x}_\star)\right).$$

4. What does it mean for a classifier to be calibrated? Briefly explain the idea behind a reliability diagram and the expected calibration error.