Bayesian Machine Learning

Georgi Tancev

April 11, 2020

Introduction

- Frequentist statistics is strictly valid for the limiting case when $|N| \gg |\theta|$ and suffers from several limitations.
  - No uncertainty information in predictions (we usually only model the expectation).
  - No possibility of data generation (no generative models).
  - Regularization only via cross-validation (maximum a posteriori estimates).
  - No inclusion of prior information possible.
  - Online learning not available.
  - Missing value and data generation treatment not possible.
- Bayesian statistics is a theory in the field of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief (e.g. prior knowledge) in an event.
- This differs from a number of other interpretations of probability, such as the frequentist interpretation that views probability as the limit of the relative frequency of an event after many trials.
- Even though Bayesian statistics offers solutions to the mentioned problems, it also comes with its own downsides.
  - The subject is mathematically more challenging that classical approaches (just because it's less taught).
  - The assumptions in (hierarchical) models that have to be made (e.g. priors) need some experience.

Bayesian Statistics

- Bayesian statistics revolves around the usage of the Bayes' theorem, as well as sum and product rules of probability.
  - $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y \in Y} p(x|y)p(y)} = \frac{p(x|y)p(y)}{Z}$
- In principle, any parametric model such as linear regression, logistic regression, or neural network can be formulated as corresponding Bayesian version with training set (y_t, X_t) by stating a prior $p(\beta)$ over the parameters
  - $p(\beta|y_t, X_t) = \frac{p(y_t, X_t|\beta)p(\beta)}{p(y_t, X_t)} = \frac{p(y_t, X_t|\beta)p(\beta)}{\int p(y_t, X_t|\beta)p(\beta)d\beta}$
- which can be computed using sampling, i.e. Markov chain Monte Carlo, or variational methods, in which the latter is (much) faster but less accurate, since exact inference is infeasible due to the intractability of the normalization integral in the denominator (with the exception of Bayesian linear regression using a conjugate prior over the parameters).
- A prediction for a new sample x is performed by marginalizing the parameter β from the conditional posterior distribution
  - $p(y = 1|x, y_t, X_t) = \int p(y = 1|x, \beta)p(\beta|y_t, X_t)d\beta$
- which is essentially a convolution of two distributions that can be calculated by sampling or in case of variation inference using numerical methods (due to mean field approximation), as once again this integral might be intractable.

Metropolis Algorithm

- Drawing samples from a multi-dimensional (possibly not normalized) probability distribution $p(z)$, e.g. $p(y|x) \propto p(x|y)p(y)$, is challenging.
- The idea of sampling algorithms is to draw samples iteratively from a simpler proposal distribution $\hat{p}(z)$, e.g. a factorized one-dimensional Gaussian.
- One such Monte Carlo method is known as Metrolpolis algorithm, which generates a Markov chain and eventually converges to the desired distribution $p(z)$.
  1. Initialize a starting point z_0 and a maximum length t_{max}.
  2. At each cycle t of the algorithm with current sample z_t, a candidate $\hat{z}$ is drawn from the proposal distribution $\hat{p}(z)$.
  3. The sample $\hat{z}$ is accepted with probability $A(\hat{z}, z_t) = min(1, \frac{p(\hat{z})}{p(z_t)})$, which can be achieved by choosing a random number $u \sim U(0, 1)$ and accepting the sample if $A(\hat{z}, z_t) > u$. If the step from z_t to $\hat{z}$ causes an increase in the value of $p(z)$, then the candidate point is certain to be kept.
  4. If the candidate sample is accepted, then $z_{t+1} = \hat{z}$, otherwise the candidate point $\hat{z}$ is discarded and z_{t+1} is set to z_t, and another sample is drawn.
  5. The process is repeated until t_{max} is reached, and the first half of the chain is discarded.
- In practice, due to increased computational power, several chains are simulated in parallel to check for convergence and mixing.

Markov Chains

- To check for convergence and mixing of Markov chains, several metrics have been developed such as
- effective sample size $\hat{N}_{eff}$,
  - that is a diagnostic computed as $\hat{N}_{eff} = \frac{MN}{\hat{\tau}}$ with $\hat{\tau} = -1 + 2 \sum_{t'=0}^{K} \hat{P}_{t'}$, where M is the number of chains, N the number of draws, $\hat{\rho}_t$ is the estimated autocorrelation at lag t, and K is the last integer for which $\hat{P}_K = \hat{\rho}_K + \hat{\rho}_{K+1}$ is still positive;
- rank normalized $\hat{R}$,
  - which is a diagnostic that tests for lack of convergence by comparing the variance between multiple chains to the variance within each chain. If convergence has been achieved, the between-chain and within-chain variances should be identical. To be most effective in detecting evidence for non-convergence, each chain should have been initialized to starting values that are dispersed relative to the target distribution.

    The diagnostic is computed by $\hat{R} = \frac{\hat{V}}{W}$ where W is the within-chain variance and $\hat{V}$ is the posterior variance estimate for the pooled rank-traces. This is the potential scale reduction factor, which converges to unity when each of the traces is a sample from the target posterior. Values greater than one indicate that one or more chains have not yet converged;
- Markov chain standard error (MCSE),
  - which is the uncertainty added from the simulation due to small sample sizes.
- In addition, probability distributions are characterized by the so-called highest posterior density (HPD),
  - that is the minimum width (narrowest) Bayesian credible interval (BCI), within which an unobserved parameter value falls with a particular probability.

Bayesian Logistic Regression

- Logistic regression is a powerful model that allows to analyze how a set of features affects some binary target label.
  - The posterior distribution gives the interval estimates for each weight of the model. This is essential in data analysis to not only provide a good model but also an uncertainty estimate of the conclusions.
- We have a collection of wine bottles; we want to develop a classification model for one specific type of wine which is often imitated due to its high price. Furthermore, we want to act only when we're (not) confident enough that it's (not) an imitation.
- To characterize a wine, we use it's chemical composition with respect to flavanoids and proline.

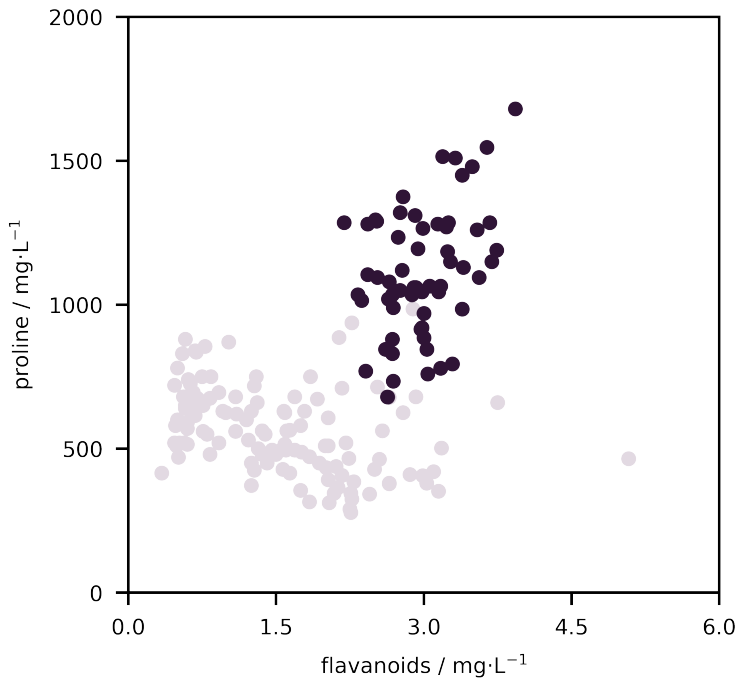

Figure: Scatter plot of the different wines (relevant one in dark shade).

Inference

- A linear model $p(y = 1|x, \beta) = \frac{1}{1+e^{-(\beta^T x+\beta_0)}}$ with priors $\beta_i \sim \mathcal{N}(0, 10^6)\ \forall\ i$ is inferred by sampling using the Python library PyMC. As the variables are not standardized before analysis, the magnitude of the parameters has no meaning.

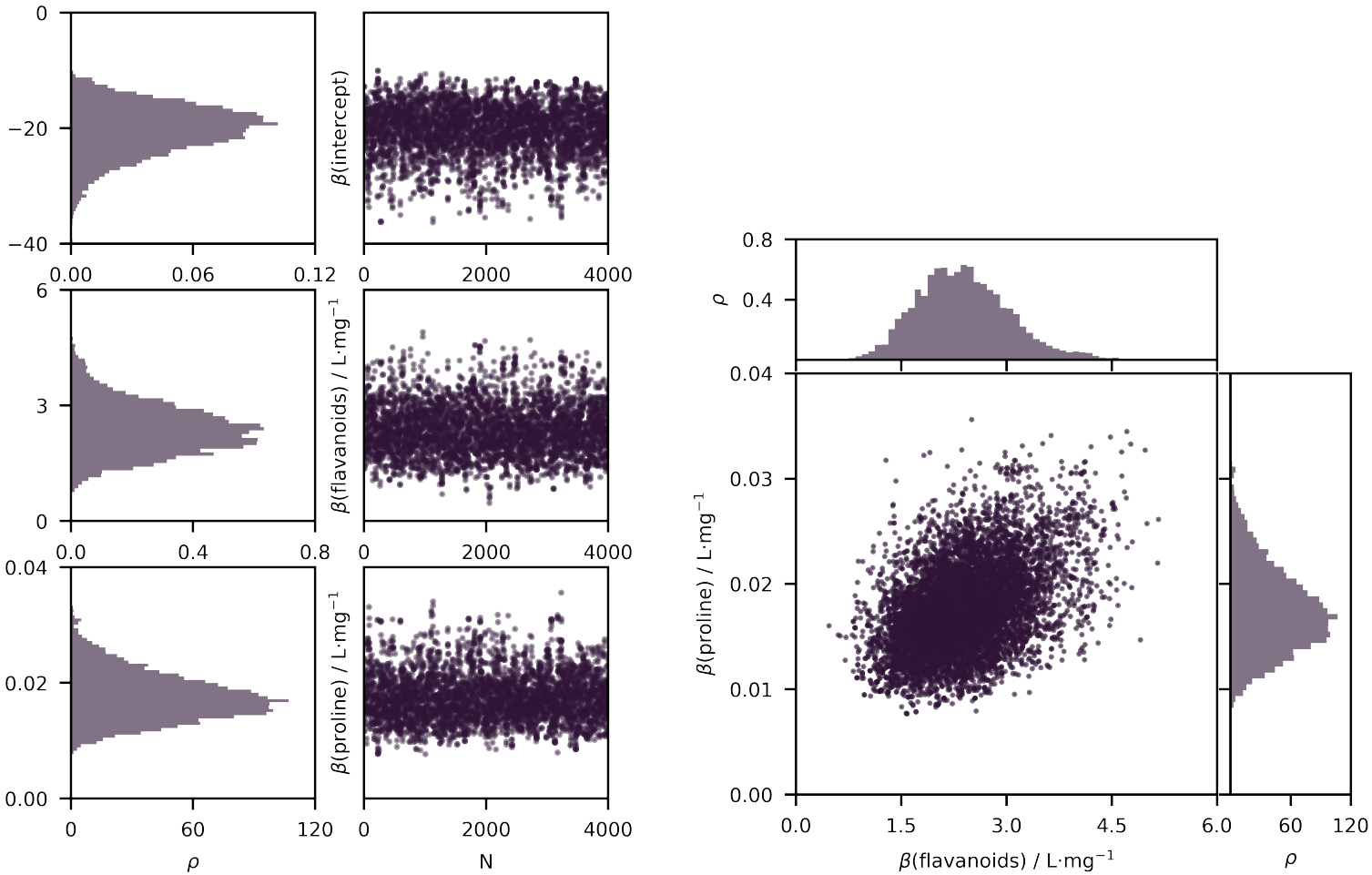

Figure: Markov chain of the inference procedure with (joint) distribution of parameters. Trace shows good mixing and convergence (at least visually).

Model Checking

- A prediction for a new sample x corresponds to $p(y = 1|x, y_t, X_t)$.

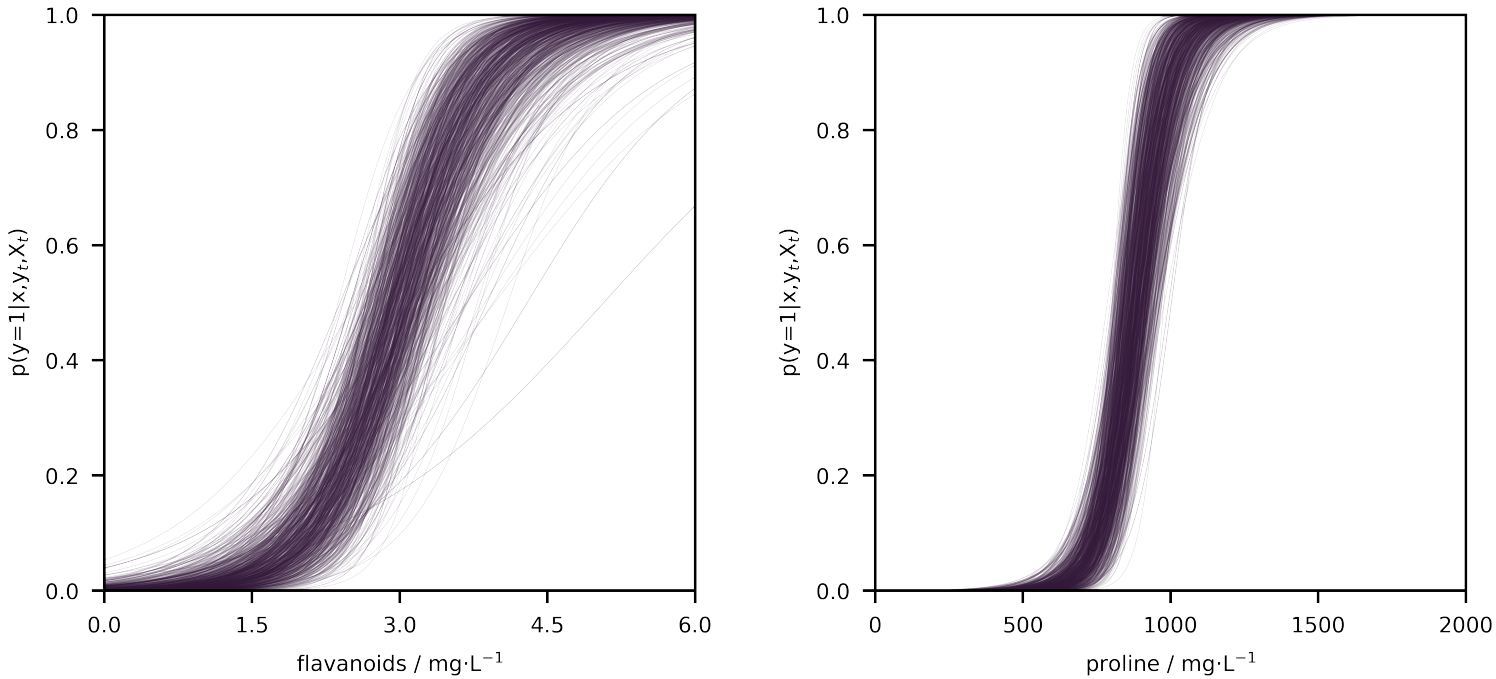

Figure: Uncertainty in predictions visualized with opacity and bandwidth. Compared to proline, classification according to flavanoids contains more uncertainty.

Table: Summary of model and inference. The chains have converged as indicated by the metrics.

parameter	μ	σ	$q_{HPD,3\%}$	$q_{HPD,97\%}$	μ_{MCSE}	σ_{MCSE}	$\hat{N}_{eff}$	$\hat{R}$
β(intercept)	-20.655	4.658	-28.548	-12.667	0.153	0.108	862.0	1.0
β(flavanoids)	2.409	0.697	1.248	3.604	0.018	0.013	1365.0	1.0
β(proline)	0.018	0.004	0.010	0.026	0.000	0.000	912.0	1.0

Prediction

- We have two new bottles of wine. How sure can we be that it's (or it's not) the wine we're looking for?
- The probability $p(y = 1|x, y_t, X_t)$ has to be computed.
  - The first bottle has a 94% credible interval of [0.90, 1.00] with $q_{0.5} = 0.97$.
  - The second bottle has a 94% credible interval of [0.50, 1.00] with $q_{0.5} = 0.85$, hence we're less confident in this case.

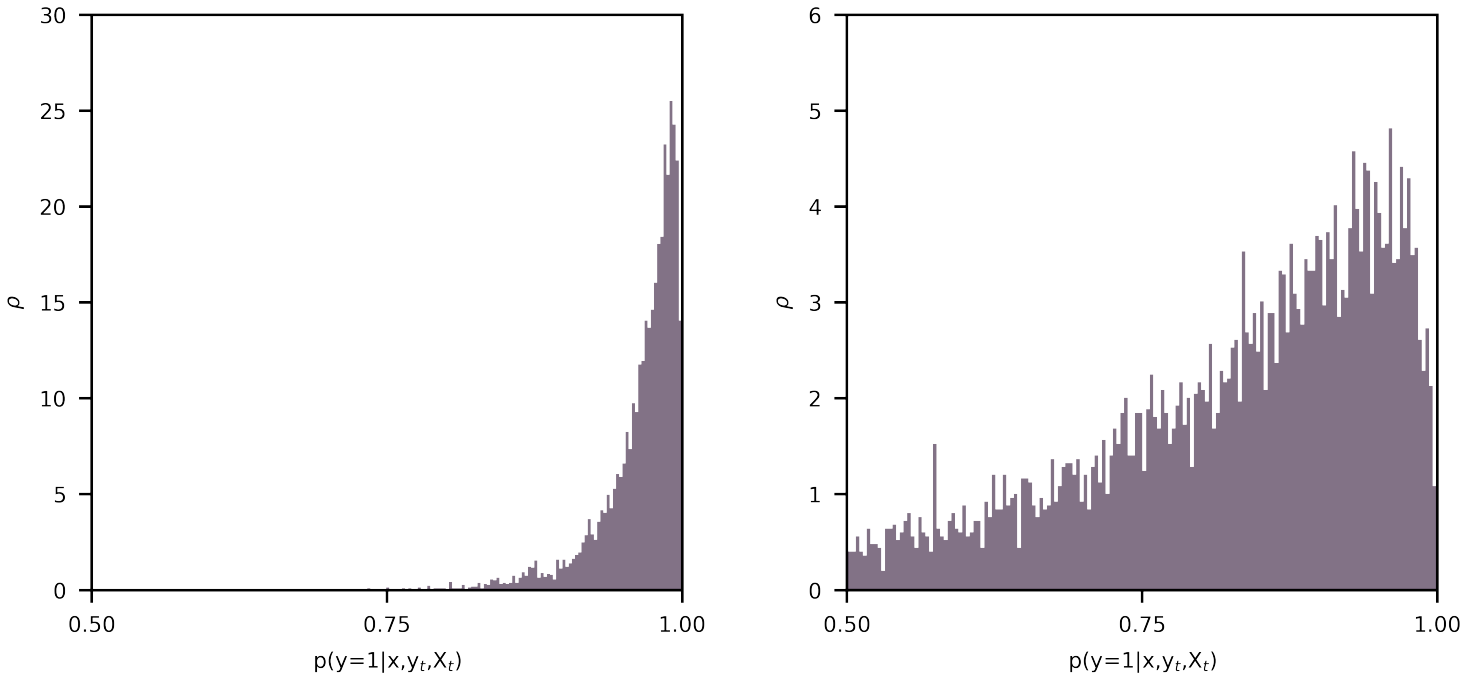

Figure: Uncertainty in predictions as bandwidth with respect to posterior probability for two new sample.