QBUS6810
Statistical Learning and Data Mining

Semester 2, 2021

Week 11 Tutorial Exercises

When studying these exercises, please keep in mind that they are about problem-solving techniques for statistical learning. In general, they're not about particular distributions or learning algorithms.

Question 1

Let $Y_1, Y_2, \ldots, Y_n \sim \text{Poisson}(\lambda)$. Recall that the Poisson distribution has probability mass function

$$p(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}.$$

(a) Write down the likelihood for a sample $y_1, \ldots, y_n$.

(b) Derive a simple expression for the log-likelihood.

(c) Let the objective function for optimisation be the negative log-likelihood. Find the critical point of the cost function.

(d) Show that the critical point is the MLE.

(e) You can create many additional exercises of this type by picking any simple statistical distribution and answering the same questions.

Question 2

In addition to being good practice, this exercise derives results that will be very useful later.

Consider the model $Y_1, Y_2, \ldots, Y_n \sim \text{Bernoulli}\Big(\sigma(\beta)\Big)$, where $\beta \in \mathbb{R}$ is a parameter and σ is the sigmoid function

$$\sigma(\beta) = \frac{1}{1 + \exp(-\beta)}.$$

You can think of this model as a logistic regression that only has the intercept.

Following the lecture, the optimisation problem for estimating this model is

$$\underset{\beta}{\text{minimise}} \quad \left\{\sum_{i=1}^{n} -y_i \log\Big(\sigma(\beta)\Big) - (1-y_i)\log\Big(1-\sigma(\beta)\Big)\right\},$$

(a) Differentiate $\sigma(\beta)$.

(b) Show that $\sigma'(\beta) = \sigma(\beta)(1-\sigma(\beta))$.

(c) Find the derivative of $J(\beta)$ using the chain rule and the previous result.

(d) Find the critical point of $J(\beta)$.

(e) What is the second derivative of the cost function? Show that the objective function is convex.

Question 3

Suppport vector machines (SVMs) were a major development in machine learning in the mid-1990s due to their state-of-art performance and novelty at the time. Since then, researchers have discovered that support vector machines can be reformulated as regularised estimation, establishing a deep connection to classical methods such as logistic regression.

In *suppport vector classification* (SVC), we consider a binary classification problem and encode the response as $y \in \{-1, 1\}$. The method is based on the linear *decision function*

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

and classification rule

$$\widehat{y} = \text{sign}\Big(f(\boldsymbol{x})\Big),$$

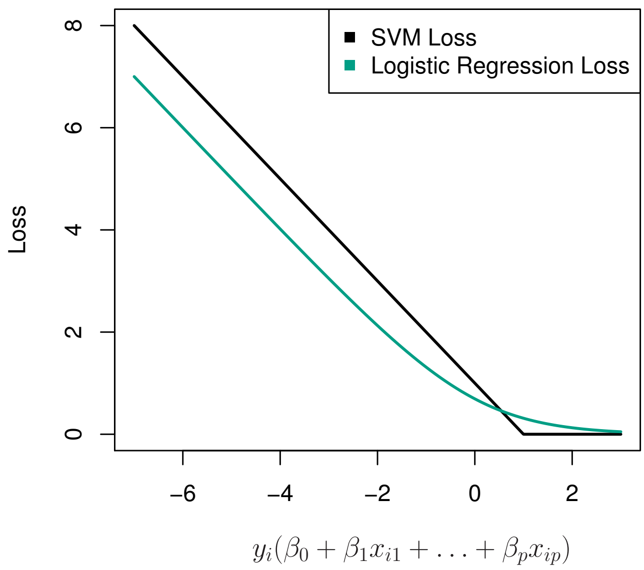
which means that $\widehat{y} = 1$ if $f(\boldsymbol{x}) > 0$ and $\widehat{y} = -1$ if $f(\boldsymbol{x}) < 0$.

The set $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$ is the *decision boundary.* Thus, we can view $|f(\boldsymbol{x})|$ as a measure of the learning algorithm's confidence that the observation is correctly classified.

The support vector classifier learns the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ by regularised empirical risk minimisation based on the *hinge loss*

$$L\Big(y, f(\boldsymbol{x})\Big) = \max\Big\{0, 1 - yf(\boldsymbol{x})\Big\}.$$

This figure from the ISL textbook plots the hinge loss and the cross-entropy loss (negative log-likelihood loss) for $y = 1$. The figure calls the latter the logistic regression loss because in this formulation, the prediction $f(\boldsymbol{x})$ in the loss function $L(y, f(\boldsymbol{x}))$ is a prediction for the logit of the probability.



(a) Write down the learning rule for a support vector classifier based on ℓ_2 regularisation.

(b) Consider the term $y_i f(\boldsymbol{x}_i)$ from the hinge loss. What is the classification when $y_i f(\boldsymbol{x}_i) > 0$ compared to $y_i f(\boldsymbol{x}_i) < 0$?

(c) Intepret the hinge loss function by considering the following cases:

1. $y f(\boldsymbol{x}) > 1$
2. $0 < y f(\boldsymbol{x}) < 1$
3. $y f(\boldsymbol{x}) < 0$

(d) The observations that satisfy $y_i f(\boldsymbol{x}_i) \leq 1$ are called the *support vectors*. Why do the support vectors have special relevance in this method?

(e) Interpret the figure from the beginning of the exercise. Compare the hinge and logistic regression loss functions to the zero-one loss and to each other. What do we learn about loss functions for binary classification?

Question 4

This exercise is left as homework as there is probably not enough time cover it in the tutorial.

Consider the generalised linear model

$$
\begin{aligned}
Y_i | X_i = x_i &\sim \text{Poisson}(\mu_i) \\
\log(\mu_i) &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.
\end{aligned}
$$

Recall that the Poisson distribution has probability mass function

$$
p(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}.
$$

(a) Write down the likelihood function for a sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$.

(b) Identify some of the differences and similarities between this exercise and Question 1.

(c) Derive an expression for the log-likelihood.

(d) Write down the optimisation problem for ℓ_1-regularised estimation of this model.