ELEC70122: Machine Learning for Actionable Decision-Making in Healthcare

Problem Set 1

Due date: Thurs 6 Feb, 2025
Total Marks: 30

Submission Instructions

Upload a pdf of you solutions to the following problems to blackboard assignments for marking. Late submissions will be penalized 10% per day late and will NOT be marked if delayed by more than 3 days. Should you have problems with uploading, please contact the TAs listed on blackboard.

---

1 Bayesian Decision Making and Linear Regression

1. Consider the example of a doctor's diagnosis for you. The doctor has good news and bad news. The bad news is that you tested positive for a serious disease and that the test is 90% accurate (i.e. the probability of testing positive given that you have the disease is 0.90, as is the probability of testing negative given that you don't have the disease). The good news is that it is a rare disease striking only 10 in 10000 people. Should you be worried? Explain your results through calculation. (3 marks)

2. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a set of d features for n patients e.g. heart rate, blood pressure, cholesterol etc. Let $\mathbf{y} \in \mathbb{R}^n$ denote a corresponding set of n labels. Assume the likelihood of the data we observe is Gaussian:

$$p(\mathbf{y}|\mathbf{X}, \theta, \mathbf{\Sigma}) = |2\pi\mathbf{\Sigma}|^{-1/2} e^{-1/2(\mathbf{y}-\mathbf{X}\theta)^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{y}-\mathbf{X}\theta)}. \quad (1)$$

   Assume that the prior of θ is also Gaussian:

$$p(\theta) = |2\pi\mathbf{\Delta}|^{-1/2} e^{-1/2\theta^T \mathbf{\Delta}^{-\mathbf{1}}\theta} \quad (2)$$

   Using Bayes rule and completing squares, derive an expression for the posterior distribution of θ. In this part, assume that the covariance $\mathbf{\Sigma}$ is given. State clearly what the mean and variance of the posterior are. Also, state the conditions under which the posterior mean would be equivalent to the ridge and maximum likelihood estimators. (4 marks)

3. Comment on the reasons for why one would want to quantify uncertainty for decision-making in healthcare. Refer to what makes healthcare unique and provide specific examples of scenarios where this may be useful. (3 marks)

2 Electronic Health Records and Risk Stratification

1. Describe why risk stratification is important for clinical applications? (2 marks)

2. Supervised topic models enable clinical researchers to identify interpretable co-occurrence patterns in count data relevant to diagnostics. However, traditional supervised Latent Dirichlet Allocation (sLDA) models face two key challenges. First, when documents contain significantly more words than labels, the influence of the labels becomes negligible. Second, due to conditional independence assumptions in the graphical model, supervised labels have limited impact on the learned topic-word distributions, often resulting in poor predictive performance on held-out data. Describe in words how you might resolve these issues? (5 marks)

3. Comment on some limitations of switching state-space models in the context of ICU intervention onset and what you would have done differently or how you would resolve these. (3 marks)

3 Survival Analysis and Learning to Defer

1. How are data in healthcare typically censored and why? (2 marks)

2. Why can't we use MSE for survival analysis with censored data? What is the alternative? (3 marks)

3. The learning-to-defer framework considers relying on human expertise in either uncertainty or where a machine learning model makes mistakes. Can you think of healthcare contexts where this may not be suitable? What are the alternatives in these cases? (5 marks)
