ELEC70122: Machine Learning for Actionable Decision-Making in Healthcare

Problem Set 1 Solutions

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

Submission Instructions

Upload a PDF of your 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 10,000 people. Should you be worried? Explain your results through calculation. (3 marks)

   Solution: A doctor informs you that you tested positive for a serious disease. The test is 90% accurate, meaning:

   - The probability of testing positive given that you have the disease is $P(T^+|D) = 0.90$.
   - The probability of testing negative given that you don't have the disease is $P(T^-|\neg D) = 0.90$.
   - The disease is rare, affecting only 10 in 10,000 people: $P(D) = \frac{10}{10000} = 0.001$.
   - The probability of not having the disease is $P(\neg D) = 1 - 0.001 = 0.999$.

   We want to determine whether you should be worried by calculating the probability that you actually have the disease given that you tested positive, $P(D|T^+)$, using Bayes' Theorem:

   $$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)} \tag{1}$$

   where

   $$P(T^+) = P(T^+|D)P(D) + P(T^+|\neg D)P(\neg D) \tag{2}$$

   Since the probability of a false positive is $P(T^+|\neg D) = 1 - 0.90 = 0.10$, we substitute the values:

   $$\begin{aligned} P(T^+) &= (0.90 \times 0.001) + (0.10 \times 0.999) \\ &= 0.0009 + 0.0999 \\ &= 0.1008 \end{aligned}$$

   Now, we compute $P(D|T^+)$:

$$
\begin{aligned}
P(D|T^+) &= \frac{0.90 \times 0.001}{0.1008} \\
&= \frac{0.0009}{0.1008} \\
&\approx 0.0089 \text{ (or } 0.89\%)
\end{aligned}
$$

Despite testing positive, the probability that you actually have the disease is only **0.89%**. This is because the disease is very rare, and even an accurate test produces many false positives. While a follow-up test is advisable, this result alone is not a strong cause for concern.

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

$$p(y|X, \theta, \Sigma) = |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}(y-X\theta)^T \Sigma^{-1}(y-X\theta)}. \tag{3}$$

Assume that the prior of θ is also Gaussian:

$$p(\theta) = |2\pi\Delta|^{-1/2} e^{-\frac{1}{2}\theta^T \Delta^{-1}\theta}. \tag{4}$$

Using Bayes' rule and completing squares, derive an expression for the posterior distribution of θ. In this part, assume that the covariance Σ 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)

Solution: Using Bayes' theorem:

$$p(\theta|y, X) \propto p(y|X, \theta)p(\theta). \tag{5}$$

Expanding the exponents:

$$
\begin{aligned}
\log p(\theta|y, X) &\propto -\frac{1}{2}(y - X\theta)^T \Sigma^{-1}(y - X\theta) - \frac{1}{2}\theta^T \Delta^{-1}\theta. \\
&= -\frac{1}{2}y^T \Sigma^{-1} y + \theta^T X^T \Sigma^{-1} y - \frac{1}{2}\theta^T X^T \Sigma^{-1} X\theta - \frac{1}{2}\theta^T \Delta^{-1}\theta.
\end{aligned}
$$

Completing the square, we rewrite:

$$\log p(\theta|y, X) \propto -\frac{1}{2}\left(\theta^T (X^T \Sigma^{-1} X + \Delta^{-1})\theta - 2\theta^T X^T \Sigma^{-1} y\right).$$

Recognizing this as a Gaussian form, we identify:

$$
\begin{aligned}
\mu_\theta &= (X^T \Sigma^{-1} X + \Delta^{-1})^{-1} X^T \Sigma^{-1} y, \\
\Sigma_\theta &= (X^T \Sigma^{-1} X + \Delta^{-1})^{-1}.
\end{aligned}
$$

Thus, the posterior distribution is:

$$p(\theta|y, X) = \mathcal{N}(\mu_\theta, \Sigma_\theta). \tag{6}$$

Special cases:

- If the prior is uninformative ($\Delta^{-1} \to 0$), then the posterior mean simplifies to the Maximum Likelihood Estimator (MLE):
$$\hat{\theta}_{\text{MLE}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y. \tag{7}$$
- If the prior is Gaussian with $\Delta = \lambda I$ (ridge regression prior), then the posterior mean simplifies to the Ridge Regression estimator:
$$\hat{\theta}_{\text{MAP}} = (X^T X + \lambda I)^{-1} X^T y. \tag{8}$$

3. Comment on the reasons 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)

Solution:

One fundamental challenge in healthcare decision-making is that data is often only available from sick patients, leading to selection bias and uncertainty in predictions for healthy individuals or those at early stages of disease. Most medical data is collected from patients who seek treatment, meaning that we primarily observe data from individuals who are already diagnosed or symptomatic. This introduces several issues:

a) Lack of representation of healthy individuals: Models trained on hospital data may struggle to predict disease risk in the general population.
b) Underestimation of disease prevalence: Since asymptomatic or undiagnosed individuals are not included, a model may **fail to recognize early-stage disease patterns.
c) Overconfident predictions: If a model is trained only on severe cases, it might be **overly confident** in diagnosing mild or borderline conditions.

Because data is missing for healthy or undiagnosed individuals, **model predictions outside the observed population come with high uncertainty**. This is especially critical for preventative medicine and screening tests. E.g. - In cancer screening, AI models trained on biopsy confirmed cancer cases may be biased toward identifying late-stage cancers, missing subtle early-stage indicators. Given this incomplete data problem, it is crucial that healthcare AI models expresss uncertainty for predicting unobserved groups, identify when training data does not generalize to new patients and defer to human experts when confidence is low

E.g. ICU mortality prediction, if a model has only been trained on critically ill patients, it should **express high uncertainty** when asked to predict outcomes for a patient with mild symptoms. Because healthcare datasets primarily come from sick patients, **uncertainty quantification is essential** for ensuring reliable predictions, particularly in **preventive medicine, early diagnosis, and risk assessment for asymptomatic individuals**. Recognizing the limitations of observed data helps prevent **biased and overconfident decision-making** in clinical practice.

2 Electronic Health Records and Risk Stratification

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

   Solution: Risk stratification is crucial in healthcare applications because it helps to prioritize and tailor interventions based on the likelihood of adverse outcomes for patients. By assessing risk factors, healthcare providers can identify patients who are at higher risk for complications or poor health outcomes, allowing for more proactive and personalized care.

   Targeted Interventions: Risk stratification helps healthcare professionals focus resources on patients who need them most. By identifying high-risk individuals early, interventions can be initiated promptly, potentially preventing or reducing the severity of health issues.

   Improved Patient Outcomes: By understanding a patient's risk level, healthcare providers can deliver more appropriate treatments and follow-ups, leading to better health outcomes. For example, early identification of high-risk patients with chronic conditions can prevent hospital readmissions and complications.

   Efficient Resource Allocation: Healthcare resources are often limited, and risk stratification helps to allocate them more efficiently. This ensures that higher-risk patients receive timely attention while lower-risk patients may be monitored with less intensity.

   Cost Reduction: By preventing avoidable complications, hospitalizations, and interventions, risk stratification can lead to significant cost savings for both healthcare providers and patients.

   Personalized Care Plans: It allows for more personalized care, where the treatment approach is based on the specific risks a patient faces. This enhances the effectiveness of treatments and can improve the patient experience.

   Predictive Modeling: Advanced risk stratification often relies on predictive models, which can analyze large datasets to forecast future health risks. This helps healthcare systems prepare for and respond to potential healthcare crises more effectively.

   Public Health Impact: On a larger scale, risk stratification can improve population health by identifying at-risk groups for various conditions (e.g., cardiovascular diseases, diabetes) and guiding public health initiatives accordingly.

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:

(a) When documents contain significantly more words than labels, the influence of the labels becomes negligible.

(b) 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)

Solution: One potential solution is to modify the model to better incorporate label information, especially in cases where the number of labels is significantly smaller than the number of words (e.g., in clinical applications where labels might represent diseases or conditions while the data consists of a large set of clinical terms or notes). This could be done by adjusting the priors to ensure that the labels are still influential even when their count is smaller. For example, adding a hierarchical structure could allow labels to influence the topic distributions more strongly, even when they are fewer in number. Alternatively, using regularization techniques could help by introducing a penalty that encourages the model to better incorporate label information.

The conditional independence assumption in the standard sLDA model can limit the impact of supervised labels on the topic-word distributions. One approach to overcoming this is to relax this assumption by incorporating dependency structures that allow for a more direct interaction between the labels and the topic-word distributions. This could be achieved through graphical models that allow for richer dependencies or by integrating neural network-based models that use supervised labels more effectively. For instance, neural topic models can be employed to learn joint representations of topics and labels, which can better capture the relationship between the clinical labels and the topics in the text data. Additionally, we can consider incorporating auxiliary information (such as metadata or patient history) alongside the labels to influence the topic distributions more effectively, ensuring that the supervised information directly informs the topics.

A more direct approach might be to integrate supervised signals into the topic generation process. This could be done by conditioning topic distributions on both the words in the document and the supervised labels, either by modifying the Dirichlet prior to be label-dependent or by using variational inference methods that more explicitly incorporate label information into the topic modeling process. This would strengthen the influence of the labels on the topics generated, improving the model's predictive performance, especially on held-out data.

Another approach could be to integrate deep learning techniques, such as a combination of LDA and neural networks, where the neural network can learn a more flexible mapping between the labels and the topic distributions. This can help overcome the strict assumptions in traditional sLDA, allowing for more nuanced and dynamic relationships between labels and topics, leading to better predictive performance on new, unseen data.

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)
Solution: Switching state-space models (SSMs) are often used in healthcare settings, including ICU intervention onset, to model temporal dynamics and capture transitions between different states (e.g., patient health states). While these models have some advantages, such as their ability to represent different regimes or phases of patient health, they also have certain limitations, especially when applied to complex and dynamic ICU environments. Here are some of the limitations, followed by potential approaches to resolve or address them:

1. **Modeling Non-Markovian Transitions:** Traditional switching state-space models often assume Markovian dynamics, where the future state depends only on the current state and not on the history of past states. However, in ICU settings, the progression of a patient's condition is influenced by a sequence of events or interventions over time, which can make the assumption of memoryless transitions unrealistic. For example, the patient's response to a specific treatment could depend on previous treatments or the sequence of interventions.

To address this, I would propose extending the model to **non-Markovian state-space models** or incorporating **memory-based structures** such as recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks. These models can better capture long-term dependencies and allow for more accurate predictions of intervention needs.

2. **Oversimplified Transition Dynamics:** Switching state-space models generally use pre-defined transition rules (e.g., between health states) which may not accurately reflect the complex and non-linear nature

of patient trajectories in the ICU. The transitions might oversimplify the actual dynamics between different states (e.g., between stable and critical states).

One way to resolve this is by *incorporating machine learning techniques*, like reinforcement learning or causal inference models, to learn the optimal transitions from data rather than relying on simplistic rules. These methods can capture more complex relationships between states, helping to predict when an intervention should occur based on the history and real-time data.

3. **Data Sparsity and Incomplete Data**: ICU patient data is often sparse and can be incomplete (e.g., missing measurements, unobserved confounders, or irregular monitoring). State-space models, particularly when used in their traditional form, may struggle to handle such issues effectively, potentially leading to incorrect state estimates or inaccurate predictions of intervention onset.

To address this, I would use *imputation techniques* to handle missing data, such as GPs or autoencoders, which can help infer missing observations based on the available data. Moreover, integrating multimodal data sources (such as patient history, lab results, and real-time monitoring) could provide a more complete picture of the patient's state, reducing the reliance on any one data source and improving model accuracy.

This list is not exhaustive and other possibilities exist. Use your discretion

3 Survival Analysis and Learning to Defer

1. How are data in healthcare typically censored and why? (2 marks)
   Solution: There are three censoring types but usually healthcare data is right censored. Right censoring when we know that an event of interest (e.g., death, disease progression, or treatment failure) has not happened by the end of the observation period, but we don't know when it will happen. In healthcare, we still have to take these probabilities into account when making decisions. E.g. Just because a patient hasnt died in the ICU, does not mean that the patient is unlikely to experience death after they leave the hospital as a result of some side effects of treatment or otherwise.

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

   Solution: The Mean Squared Error (MSE) is not suitable for survival analysis with censored data because it does not appropriately account for the censoring process, which can lead to biased and misleading results.

   Censoring Issue. MSE assumes that the prediction error (difference between observed and predicted values) is calculated for every observation, without any consideration for whether the data point is censored or not. When a subject is censored, the true event time is unknown, so using MSE would involve comparing the predicted event time to an unknown value, which introduces bias and error.

   Ignoring Censoring Information. MSE treats the observed survival times as fixed outcomes, which is not accurate in the presence of censoring. The censored observations provide partial information, and their contribution to the error term should not be treated the same as fully observed data. If MSE is used, censored data points are effectively treated as if the event occurred at the observed time, which is misleading and can significantly distort the analysis.

   Inability to Handle the Time-to-Event Nature. Survival analysis models focus on estimating the time until the event occurs, not just the event outcome itself. MSE is designed for regression tasks where predictions are compared to fixed target values (e.g., predicting an exact value for continuous outcomes). Survival analysis, however, deals with probabilistic events over time, where we are interested in estimating the risk or hazard of the event occurring at a certain time, considering the time-varying nature of the data.

   Alternatives such as Cox Proportional Hazards model (with partial likelihood) or non-parametric methods like the Kaplan-Meier estimator are designed specifically to handle censored data by making use of the available information without assuming that the event has occurred for censored observations.

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)

   Solution: The learning-to-defer framework in healthcare, which involves relying on human expertise when machine learning models are uncertain or likely to make mistakes, is effective in many scenarios but may not be suitable in some contexts. In emergency situations, such as ICU care or during critical events like cardiac arrest, deferring to human experts could cause delays in life-saving decisions, potentially leading to worse patient outcomes. Similarly, in rural healthcare or during night shifts, when specialists may not be available, waiting for human input is impractical. For chronic disease management, where continuous monitoring and frequent intervention adjustments are needed, deferring decisions to clinicians could be inefficient and burdensome. Large-scale public health efforts, like predicting infectious disease spread, require rapid, autonomous decisions that a human expert cannot handle at the scale necessary. Additionally, relying on human oversight for each uncertainty or mistake in primary care could contribute to clinician burnout, overburdening healthcare providers. Alternatives to the learning-to-defer approach include autonomous decision support systems that act independently while flagging high-risk cases for human review. Hybrid systems, which collaborate with clinicians by prioritizing urgent cases, can ensure rapid decision-making without overwhelming experts. Continuous learning systems can improve over time, reducing uncertainty and the need for human intervention. Triage algorithms can automatically prioritize cases based on urgency, reserving human expertise for critical decisions. These alternatives offer efficient, scalable solutions that optimize healthcare delivery while maintaining safety, especially in time-sensitive or resource-constrained environments. Finally when the human expert is also uncertain, it may not be helpful to defer and second opinions are required