SLIDS Homework 5

Dave Zachariah

March 9, 2023

- Solution proposals are individual.
- Each solution must be reproducible by your peer. Code should be added to the appendix.
- The solution to each subproblem yield 0 to 2 points. Students are expected to attempt each problem.

1

We consider a deterministic policy $\pi(\mathbf{x})$ for predicting whether patients with covariates $\mathbf{x}$ will develop a heart disease $y \in \{0, 1\}$ within a given time period. The covariates $\mathbf{x} = [x_1 \ x_2]^\top$ we observe are age and LDL cholestorial level, respectively.

Thus in the decision process we have that

$$p^\pi(a|\mathbf{x}) = \mathbb{1}\{a = \pi(\mathbf{x})\}$$

Note that this implies that

$$\mathbb{E}^\pi_{a|\mathbf{x}}[a] \equiv \pi(\mathbf{x})$$

a) We consider the symmetric zero-one loss:

$$\ell(y, a) = \mathbb{1}\{y \neq a\}$$

Show that the risk-minimizing predictor (Lecture 5) can in this case be expressed in the following ways:

$$\begin{aligned}
\pi_\circ(\mathbf{x}) &= \underset{a \in \mathcal{Y}}{\arg\min} \ \mathbb{1}\{a = 1\}p(y = 0|\mathbf{x}) + \mathbb{1}\{a = 0\}p(y = 1|\mathbf{x}) \\
&= \underset{a \in \mathcal{Y}}{\arg\max} \ p(y = a|\mathbf{x}) \\
&= \underset{a \in \mathcal{Y}}{\arg\max} \ p(\mathbf{x}|y = a)p(y = a) \\
&= \mathbb{1}\left\{ \frac{p(\mathbf{x}|y = 1)p(y = 1)}{p(\mathbf{x}|y = 0)p(y = 0)} > 1 \right\}
\end{aligned}$$

Thus the classifier will parition covariate space $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ in some way depending on the conditional distributions.

b) Consider two classes of predictive policies

$$\Pi_{\text{quad}} = \left\{\pi : \pi(\mathbf{x}) = \mathbb{1}\{\mathbf{x}^\top \mathbf{C}\mathbf{x} + \mathbf{b}^\top \mathbf{x} > a\}\right\}$$

and

$$\Pi_{\text{lin}} = \left\{\pi : \pi(\mathbf{x}) = \mathbb{1}\{\mathbf{b}^\top \mathbf{x} > a\}\right\}$$

Suppose the conditional covariate distributions are Gaussian, that is,

$$p(\mathbf{x}|y = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

Show that the risk-minimizing predictor $\pi_\circ$ above also belongs to the quadratic policy class Π_{quad}.

Tip: Use the natural logarithm of both sides of the inequality inside $\mathbb{1}\{\cdot\}$.

2

Suppose the unknown data-generating distribution $p(\mathbf{x}, y)$ is given by

$$p(y) = \text{Ber}(y; 0.20) \qquad p(\mathbf{x}|y = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 50 \\ 140 \end{bmatrix} \; \boldsymbol{\Sigma}_0 = \begin{bmatrix} 64 & 9 \\ 9 & 64 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 60 \\ 160 \end{bmatrix} \; \boldsymbol{\Sigma}_1 = \begin{bmatrix} 64 & 49 \\ 49 & 64 \end{bmatrix}$$

a) Draw $m = 1000$ i.i.d. test samples and plot the covariate samples $\mathbf{x}$ (age and LDL cholesterol level) with colors corresponding to health ($y = 0$) and ill patients ($y = 1$).

We would like to study the performance of predictive policies of the form:

$$\pi(\mathbf{x}) = \mathbb{1}\{T(\mathbf{x}) > \tau\}$$

by considering two incommensurable risks:

$$L_0(\pi; \phi) = \mathbb{E}[a = 1|y = 0] \qquad L_1(\pi; \phi) = \mathbb{E}[a = 0|y = 1]$$

We will evaluate both risks via Monte Carlo approximation of the expectation using using $m = 10^4$ test samples.

b) For the first policy, we consider a logistic model of the odds:

$$T(\mathbf{x}; \boldsymbol{\theta}) = \frac{p_\theta(y = 1|\mathbf{x})}{p_\theta(y = 0|\mathbf{x})} = \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x})\right)$$

using some feature vector $\boldsymbol{\phi}(\mathbf{x})$.

It is computationally hard to perform empirical risk minimization to learn $\pi(\mathbf{x}; \boldsymbol{\theta})$ from n samples. Instead we consider instead using surprisal loss $\ell_\theta(\mathbf{x}, y) = -\ln p_\theta(y|\mathbf{x})$, when

- $\boldsymbol{\phi}'(\mathbf{x}) = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$, versus,

- $\phi''(\mathbf{x}) = \begin{bmatrix} 1 \\ \mathbf{x} \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$

NB: You may use any software package for logistic regression for ERM of $\widehat{\boldsymbol{\theta}}_n$.

Learn $\widehat{\boldsymbol{\theta}}_n$ and evaluate the two risks of the resulting policy $\pi(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)$. Specifically, *report the false positive versus false negative risks as a curve*

$$(L_0(\tau), 1 - L_1(\tau))$$

(approximated by $m = 10^4$ test samples). Grid τ from very high to low values, so that the policy ranges from $\pi(\mathbf{x}) \equiv 0$ (all healthy) to $\pi(\mathbf{x}) \equiv 1$ (all ill).

For each classifier, show three performance curves using $n = 10$, 100 and 1000 samples, respectively.

c) An alternative form motivated in Lecture 5 is to use a model of the likelihood ratio:

$$T(\mathbf{x}; \boldsymbol{\theta}) = \frac{p_{\theta_1}(\mathbf{x}|y=1)}{p_{\theta_0}(\mathbf{x}|y=0)}$$

Note that in if the models were well-specified, then setting

$$\tau = \frac{p(y=0)}{p(y=1)}$$

yields the risk-minimizing policy w.r.t. the missclassification error.

Show that Gaussian data models $p_{\theta_k}(\mathbf{x}|y=k) = \mathcal{N}(\mathbf{x}; \mathbf{m}_k, \mathbf{C}_k)$ results in a predictive policy with a quadratic partition of the covariate space, i.e., it belongs to Π_{quad}.

Show that if the model covariances are equal $\mathbf{C}_k \equiv \mathbf{C}$ for both outcomes, then the policy belongs to Π_{lin}.

d) Use the surprisal loss $\ell_\theta(\mathbf{x}, y) = -\ln p_\theta(\mathbf{x}|y)$ to learn the model-based classifier above using n training samples, when

- $p_\theta(\mathbf{x}|y=k) = \mathcal{N}(\mathbf{x}; \mathbf{m}_k, \mathbf{C})$, versus,
- $p_\theta(\mathbf{x}|y=k) = \mathcal{N}(\mathbf{x}; \mathbf{m}_k, \mathbf{C}_k)$

where $\boldsymbol{\theta} = (\mathbf{m}_0, \mathbf{m}_1, \mathbf{C})$ and $\boldsymbol{\theta} = (\mathbf{m}_0, \mathbf{m}_1, \mathbf{C}_0, \mathbf{C}_1)$, respectively.

Show that ERM $\widehat{\boldsymbol{\theta}}_n$ is given in closed form: For the first case it equals

$$\begin{aligned} \widehat{\mathbf{m}}_0 &= \mathbb{E}_{n_0}[\mathbf{x}] \\ \widehat{\mathbf{m}}_1 &= \mathbb{E}_{n_1}[\mathbf{x}] \\ \widehat{\mathbf{C}} &= \frac{n_0}{n} \mathbb{E}_{n_0}\left[(\mathbf{x} - \widehat{\mathbf{m}}_0)(\mathbf{x} - \widehat{\mathbf{m}}_0)^\top\right] + \frac{n_1}{n}\left[(\mathbf{x} - \widehat{\mathbf{m}}_1)(\mathbf{x} - \widehat{\mathbf{m}}_1)^\top\right] \end{aligned}$$

and for the second case

$$
\begin{aligned}
\widehat{\mathbf{m}}_0 &= \mathbb{E}_{n_0}[\mathbf{x}] \\
\widehat{\mathbf{m}}_1 &= \mathbb{E}_{n_1}[\mathbf{x}] \\
\widehat{\mathbf{C}}_0 &= \mathbb{E}_{n_0}\left[(\mathbf{x}-\widehat{\mathbf{m}}_0)(\mathbf{x}-\widehat{\mathbf{m}}_0)^\top\right] \\
\widehat{\mathbf{C}}_1 &= \mathbb{E}_{n_1}\left[(\mathbf{x}-\widehat{\mathbf{m}}_1)(\mathbf{x}-\widehat{\mathbf{m}}_1)^\top\right]
\end{aligned}
$$

Hint: To derive ERM $\widehat{\boldsymbol{\theta}}_n$, use the fact that the surprisal loss has the equivalent form (after removing constants):

$$\ell_\theta(\mathbf{x}, y=k) = \ln|\mathbf{C}_k| + \operatorname{tr}\left\{\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)(\mathbf{x}-\mathbf{m}_k)^\top\right\}$$

You may derive the ERM covariance matrix estimates using matrix derivatives:

$$\partial_{\boldsymbol{\Sigma}} \ln|\boldsymbol{\Sigma}| = \boldsymbol{\Sigma}^{-1} \qquad \partial_{\boldsymbol{\Sigma}} \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{W}\} = -\boldsymbol{\Sigma}^{-1}\mathbf{W}\boldsymbol{\Sigma}^{-1}$$

where $\mathbf{W}$ is a positive definite matrix. Note that $n = n_0 + n_1$.

e) Compare the performance curves of all four policies $\pi(\mathbf{x}; \widehat{\boldsymbol{\theta}}_n)$ (logistic regression vs. likelihood-ratio) and (linear vs. quadratic partition) for $n = 10$, 100 and 1000 samples, respectively. Remark on the pros and cons of each learned policy.