ST451 Bayesian Machine Learning
Week 6

Exercises

1. Consider a Naive Bayes classifier with binary y and a binary feature x . In other words assume that $y, x|y=0, x|y=1$ are Bernoulli random variables with parameters $\theta = (\theta_y, \theta_0, \theta_1)$ respectively. Assume that the data consist of n points $D = (y_i, x_i)_{i=1}^n$.

   (a) Find the maximum likelihood estimates of $\theta_y, \theta_0, \theta_1$ based on D.

   Answer: Denote by x_{0i} the x_i' s for which $y_i = 0$ and assume there are n_0 of those. Similarly denote by x_{1i} the x_i' s for which $y_i = 1$ and assume that there are n_1 of those. Note that $n_0 + n_1 = n$ and that $\sum_i y_1 = n_1$. We also get by the assumution of the model that

   $$\pi(x|y, \theta_0, \theta_1) = \prod_{i=1}^{n_0} \theta_0^{x_{0i}} (1-\theta_0)^{1-x_{0i}} \prod_{i=1}^{n_1} \theta_1^{x_{1i}} (1-\theta_1)^{1-x_{1i}}$$

   We can now write

   $$\begin{aligned} f(D|\theta) &= \pi(y|\theta_y)\pi(x|y, \theta_0, \theta_1) \\ &= \prod_{i=1}^{n} \theta_y^{y_i} (1-\theta_y)^{1-y_i} \prod_{i=1}^{n_0} \theta_0^{x_{0i}} (1-\theta_0)^{1-x_{0i}} \prod_{i=1}^{n_1} \theta_1^{x_{1i}} (1-\theta_1)^{1-x_{1i}} \\ &= \theta_y^{\sum_i y_i} (1-\theta_y)^{n-\sum_i y_i} \theta_0^{\sum_i x_{0i}} (1-\theta_0)^{n_0 - \sum_i x_{0i}} \theta_1^{\sum_i x_{1i}} (1-\theta_1)^{n_1 - \sum_i x_{1i}} \\ \log f(D|\theta) &= \sum_{i=1}^{n} y_i \log \theta_y + \left(n - \sum_{i=1}^{n} y_i\right) \log(1-\theta_y) + \sum_{i=1}^{n_0} x_{0i} \log \theta_0 + \left(n_0 - \sum_{i=1}^{n_0} x_{0i}\right) \log(1-\theta_0) \\ &+ \sum_{i=1}^{n_1} x_{1i} \log \theta_1 + \left(n_1 - \sum_{i=1}^{n_1} x_{1i}\right) \log(1-\theta_1) \end{aligned}$$

   Standard MLE calculations yield $\hat{\theta}_y = \bar{y}$, $\hat{\theta}_0 = \frac{1}{n_0} \sum_{i=1}^{n} x_{0i}$ and $\hat{\theta}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$.

   (b) Fix θ_y to its $\hat{\theta}_y$ and assign independent Beta(α, α) priors to θ_0 and θ_1. Derive the posterior of θ_0 and θ_1 and use the posterior mean as Bayes estimator of θ_0, θ_1. Compare with the MLE of θ_0, θ_1.

   Answer: Given the likelihood derived in the previous part and the prior which is proportional to $\theta_0^{\alpha-1}(1-\theta_0)^{\alpha-1}\theta_1^{\alpha-1}(1-\theta_1)^{\alpha-1}$, we get

   $$\begin{aligned} \pi(\theta_0, \theta_1|D) &\propto \theta_0^{\sum_i x_{0i}} (1-\theta_0)^{n_0 - \sum_i x_{0i}} \theta_1^{\sum_i x_{1i}} (1-\theta_1)^{n_1 - \sum_i x_{1i}} \theta_1^{\alpha-1} (1-\theta_1)^{\alpha-1} \theta_2^{\alpha-1} (1-\theta_2)^{\alpha-1} \\ &= \theta_0^{\sum_i x_{0i} + \alpha - 1} (1-\theta_0)^{n_0 + \alpha - \sum_i x_{0i} - 1} \theta_1^{\sum_i x_{1i+\alpha-1}} (1-\theta_1)^{n_1 + \alpha - \sum_i x_{1i} - 1} \\ &= \text{Beta}(\alpha + \sum_i x_{0i}, n_0 + \alpha - \sum_i x_{0i}) \text{Beta}(\alpha + \sum_i x_{1i}, n_1 + \alpha - \sum_i x_{1i}) \end{aligned}$$

   The posterior means for θ_0 and θ_1 are $\frac{\alpha + \sum_i x_{0i}}{n_0 + 2\alpha}$ and $\frac{\alpha + \sum_i x_{1i}}{n_1 + 2\alpha}$ respectively.

2. Repeat the image processing example with an image of your choice. Find a back and white bmp image distort with noise and see if you can restore it using variational inference.

   Answer: You can us the code of the computer class directly

3. **Optional:** In the text classification exercise of the computer workshop, explore whether the predictive performance of the naive Bayes classifier can be improved further by using the NLTK library (https://www.nltk.org/) to perform tasks such as lemmatising words.