EECS E6720: Bayesian Models for Machine Learning

Columbia University, Fall 2018

Homework 1: Due Sunday, September 23, 2018 by 11:59pm

Please read these instructions to ensure you receive full credit on your homework. Submit the written portion of your homework as a *single* PDF file through Courseworks (less than 5MB). In addition to your PDF write-up, submit all code written by you in their original extensions through Courseworks (e.g., .m, .r, .py, etc.). Any coding language is acceptable, but your code should be your own. Do not submit Jupyter or other notebooks, but the original source code only. Do not wrap your files in .rar, .zip, .tar and do not submit your write-up in .doc or other file type. Your grade will be based on the contents of *one* PDF file and the original source code. Additional files will be ignored. We will not run your code, so everything you are asked to show should be put in the PDF file. Show all work for full credit.

Late submission policy: Late homeworks will have 0.1% deducted from the final grade for each minute late. *Your homework submission time will be based on the time of your* ***last*** *submission to Courseworks.* Therefore, do not re-submit after midnight on the due date unless you are confident the new submission is significantly better to overcompensate for the points lost. You can resubmit as much as you like, but each time you resubmit be sure to upload **all** files you want graded! Submission time is non-negotiable and will be based on the time you submitted your last file to Courseworks. The number of points deducted will be rounded to the nearest integer.

Problem 1. (10 points)

Your friend is on a gameshow and phones you for advice. She describes her situation as follows: There are three doors with a prize behind one of the doors and nothing behind the other two. She randomly picks one of the doors, but before opening it, the gameshow host opens one of the other two doors to show that it contains no prize. She wants to know whether she should stay with her original selection or switch doors. What is your suggestion? Calculate the relevant posterior probabilities to convince her that she should follow your advice.

Problem 2. (15 points)

Let $\pi = (\pi_1, \ldots, \pi_K)$, with $\pi_j \geq 0, \sum_j \pi_j = 1$. Let $X_i \sim \text{Multinomial}(\pi)$, i.i.d. for $i = 1, \ldots, N$. Find a conjugate prior for π and calculate its posterior distribution and identify it by name. What is the most obvious feature about the parameters of this posterior distribution?

Problem 3. (30 points)

You are given a dataset $\{x_1, \ldots, x_N\}$, where each $x \in \mathbb{N}$. You model it as i.i.d. Poisson(λ). Since you don't know λ, you model it as $\lambda \sim \text{Gamma}(a, b)$.

a) Using Bayes rule, calculate the posterior of λ and identify the distribution.

b) Using the posterior, calculate the predictive distribution on a new observation,

$$p(x^*|x_1, \ldots, x_n) = \int_0^\infty p(x^*|\lambda)p(\lambda|x_1, \ldots, x_N)d\lambda$$

Problem 4. (20 points)

In this problem you will use your derivations from Problem 3 to code a naive Bayes classifier for distinguishing spam from non-spam emails. The data is provided on Courseworks.

Each 54-dimensional vector x has a label y with $y = 0$ indicating "non-spam" and $y = 1$ indicating "spam". We model the nth feature vector of a spam email as

$$p(x_n|\vec{\lambda}_1, y_n = 1) = \prod_{d=1}^{54} \text{Poisson}(x_{n,d}|\lambda_{1,d}),$$

and similarly for class 0. We model the labels as $y_n \sim \text{Bernoulli}(\pi)$. Assume independent gamma priors on all $\lambda_{1,d}$ and $\lambda_{0,d}$, as in Problem 3, with $a = 1$ and $b = 1$. For the label bias assume the prior $\pi \sim \text{Beta}(e, f)$ and set $e = f = 1$.

Let (x^*, y^*) be a new test pair. The goal is to predict y^* given x^*. To do this we use the predictive distribution under the posterior of the naive Bayes classifier. That is, for possible label $y^* = y \in \{0, 1\}$ we compute

$$p(y^* = y|x^*, X, \vec{y}) \propto p(x^*|y^* = y, \{x_i : y_i = y\})p(y^* = y|\vec{y})$$

where X and $\vec{y}$ contain N training pairs of the form (x_i, y_i). This can be calculated as follows:

$$p(x^*|y^* = y, \{x_i : y_i = y\}) = \prod_{d=1}^{54} \int_0^\infty p(x^*|\lambda_{y,d})p(\lambda_{y,d}|\{x_i : y_i = y\})d\lambda$$

The results from Problem 3 can be directly applied here. Also, as discussed in the notes

$$p(y^* = y|\vec{y}) = \int_0^1 p(y^* = y|\pi)p(\pi|\vec{y})d\pi$$

which has the solutions $p(y^* = 1|\vec{y}) = \dfrac{e + \sum_n \mathbb{1}(y_n = 1)}{N + e + f}$ and $p(y^* = 0|\vec{y}) = \dfrac{f + \sum_n \mathbb{1}(y_n = 0)}{N + e + f}$.

a) Using the marginal distributions discussed above, implement this naive Bayes classifier for binary classification in your preferred language.

b) Make predictions for all data in the testing set by assigning the most probable label to each feature vector. In a 2×2 table, list the total number of spam classified as spam, non-spam classified as non-spam, as well as the off-diagonal values (i.e., a confusion matrix). Use the provided ground truth for this evaluation.

c) Pick three misclassified emails and for each email plot its features x compared with $\mathbb{E}[\vec{\lambda}_1]$ and $\mathbb{E}[\vec{\lambda}_0]$, and give the predictive probabilities for that email. Mark the 54 points along the x-axis with their names in the readme file.

d) Pick the three most ambiguous predictions, i.e., the digits whose predictive probabilities are the closest to 0.5. Show the same information for these three emails that you showed in Problem 4(c) above.