# Homework 1 SOLUTIONS (written portion)

**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.

**SOLUTION** Final answer: She should switch. (This is the famous Monty Hall problem.)

There are many different approaches to this problem which give the same answer. However, note that the problem asks "calculate the relevant posterior probabilities", which implies that you should use Bayes' rule. We can define the following random variables:

- $S_i$: 1 if she picks door $i$ and 0 otherwise
- $O_i$: 1 if hosts opens door $i$ and 0 otherwise
- $Z_i$: 1 if the prize is behind door $i$ and 0 otherwise

Since the order of the door doesn't matter, we can assume that she initially picks door 1 and the host opens door 3 (any orders will work in the same way). We want to compute the posterior $P(Z_1 = 1|S_1 = 1, O_3 = 1)$ and $P(Z_2 = 1|S_1 = 1, O_3 = 1)$. By Bayes' rule:

$$\begin{aligned} P(Z_1 = 1|S_1 = 1, O_3 = 1) &= \frac{P(S_1 = 1, O_3 = 1|Z_1 = 1)P(Z_1 = 1)}{\sum_{j\in\{0,1\}} P(S_1 = 1, O_3 = 1|Z_1 = j)P(Z_1 = j)} \\ &= \frac{P(O_3 = 1|S_1 = 1, Z_1 = 1)P(S_1 = 1|Z_1 = 1)P(Z_1 = 1)}{\sum_{j\in\{0,1\}} P(O_3 = 1|S_1 = 1, Z_1 = j)P(S_1 = 1|Z_1 = j)P(Z_1 = j)} \\ &= \frac{\frac{1}{2}\times\frac{1}{3}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}\times\frac{1}{3} + 1\times\frac{1}{3}\times\frac{1}{3}} = \frac{1}{3} \end{aligned}$$

Here we are making use of the fact that $S_1$ and $Z_1$ are independent and $O_3$ is dependent on $S_1$ and $Z_1$. (If she picks door 1 and the prize is behind door 2, the host has to pick door 3. Otherwise the host could've picked either door 2 or door 3 with equal probabilities.) We can compute $P(Z_2 = 1|S_1 = 1, O_3 = 1)$ in the similar way. However, since $P(Z_2 = 1|S_1 = 1, O_3 = 1) = P(Z_1 = 0|S_1 = 1, O_3 = 1)$, we can immediately get that it equals $\frac{2}{3}$, which leads to the conclusion that she should switch.

**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 $\pi$ and calculate its posterior distribution and identify it by name. What is the most obvious feature about the parameters of this posterior distribution?

**SOLUTION**

The likelihood given data $\mathbf{X} = \{X_i\}_{i=1}^N$:

$$p(\mathbf{X}|\pi) = \prod_{i=1}^{N} p(X_i|\pi) = \prod_{i=1}^{N} \prod_{j=1}^{K} \pi_j^{\mathbb{1}[X_i=j]}$$

According to, e.g., Wikipedia, the conjugate prior is the Dirichlet distribution

$$Dir(\alpha_1, \cdots, \alpha_K) = \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\sum_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^{K} \pi_j^{\alpha_j - 1}$$

We would also accept if $\alpha_j$ is replaced with one symbol not varying in $j$, e.g., $\alpha$, in which case the constant out front is $\Gamma(K\alpha)/\Gamma(\alpha)^K$. Multiplying these together,

$$\begin{aligned} p(\pi|\mathbf{X}) &\propto p(\mathbf{X}|\pi)p(\pi) && (1) \\ &\propto \prod_{j=1}^{K} \pi_j^{\sum_i \mathbb{1}[X_i=j]+\alpha_j-1} && (2) \end{aligned}$$

This is a Dirichlet distribution with parameters $(\alpha'_1, \ldots, \alpha'_K)$ where $\alpha'_j = \sum_i \mathbb{1}[X_i = j] + \alpha_j$. The most obvious feature is that we only need to calculate the histogram of $\mathbf{X}$. In other words, we only need to know how many times $X_j = j$, not the specific sequence of $\mathbf{X}$.

**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($\lambda$). Since you don't know $\lambda$, you model it as $\lambda \sim \text{Gamma}(a, b)$.

a) Using Bayes rule, calculate the posterior of $\lambda$ and identify the distribution.

**SOLUTION**

$$\begin{aligned} p(\lambda|x_1, \ldots, x_N) &\propto p(x_1, \ldots, x_N|\lambda)p(\lambda) && (3)\\ &\propto \prod_{i=1}^{N} p(x_i|\lambda)p(\lambda) && (4)\\ &\propto \left[\prod_{i=1}^{N} \lambda^{x_i}\mathrm{e}^{-\lambda}\right]\lambda^{a-1}\mathrm{e}^{-b\lambda} && (5)\\ &\propto \lambda^{\sum_i x_i + a - 1}\mathrm{e}^{-(b+N)\lambda} && (6)\end{aligned}$$

Finally, we can observe that this is a Gamma($a+\sum_i x_i, b+N$) distribution. It is possible that the Gamma distribution form was used where $b$ was replaced by $1/b$ above. In this case the second parameter of the posterior must be $b/(1+Nb)$.

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

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

**SOLUTION**

Let the posterior be Gamma($a', b'$).

$$\begin{aligned} p(x^*|x_{1:N}) &= \int_0^\infty \frac{\lambda^{x^*}}{x^*!}\mathrm{e}^{-\lambda}\frac{b'^{a'}}{\Gamma(a')}\lambda^{a'-1}\mathrm{e}^{-b'\lambda}d\lambda && (7)\\ &= \frac{b'^{a'}}{x^*!\Gamma(a')}\int_0^\infty \lambda^{x^*+a'-1}\mathrm{e}^{-(1+b')\lambda}d\lambda && (8)\end{aligned}$$

To solve this, it is enough to recognize that this integral is the normalizing constant of a Gamma distribution with parameters $x^* + a'$ and $1 + b'$. Therefore,

$$\int_0^\infty \lambda^{x^*+a'-1}\mathrm{e}^{-(1+b')\lambda}d\lambda = \frac{\Gamma(x^*+a')}{(1+b')^{x^*+a'}} \quad (9)$$

and

$$p(x^*|x_{1:N}) = \frac{\Gamma(x^*+a')}{x^*!\Gamma(a')}\frac{b'^{a'}}{(1+b')^{x^*+a'}} = \frac{\Gamma(x^*+a')}{x^*!\Gamma(a')}\left(\frac{b'}{1+b'}\right)^{a'}\left(\frac{1}{1+b'}\right)^{x^*} \quad (10)$$

and recall that $a' = a + \sum_{i=1}^N x_i$ and $b' = b + N$.

**Problem 4.** (20 points)

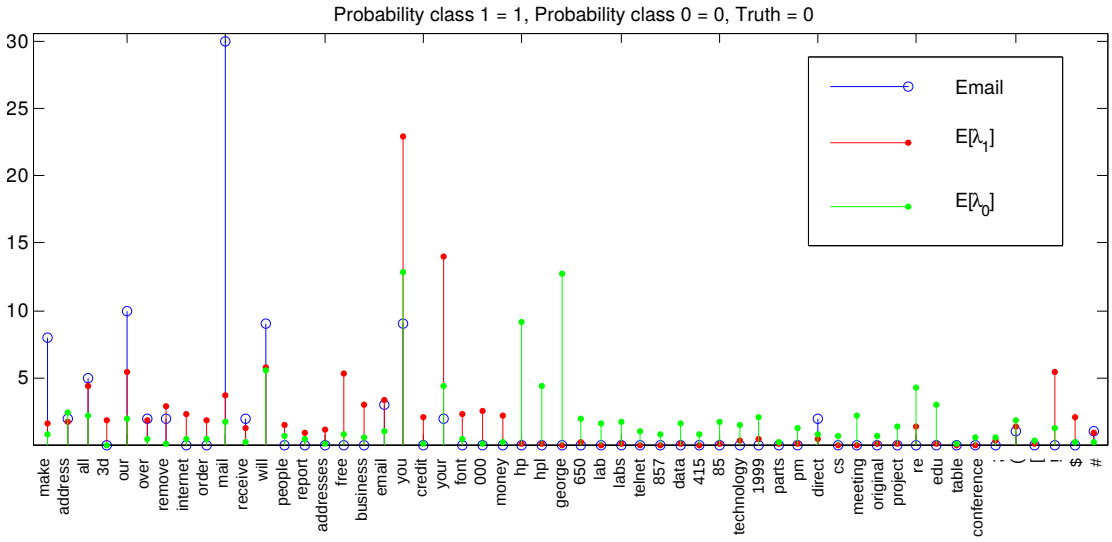
b) Make predictions for all data in the testing set by assigning the most probable label to each feature vector. In a $2 \times 2$ table, list the total number of spam classified as spam, non-spam classified as non-spam, as well as the off-diagonal values (i.e., a confusion matrix). Use the provided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

**SOLUTION** The algorithm is deterministic, so there is technically only one correct answer.

| truth↓ classify→ | spam | not spam |
|---|---|---|
| spam | 172 | 10 |
| not spam | 48 | 231 |

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.

**SOLUTION** Here is one example.




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.

**SOLUTION** Here is one example.