ML1 Homework 2

Monday, September 7, 2015
Deadline: Friday, September 18, 2015, 23:59

1 Probability distributions, likelihoods, and estimators

For these questions you will be working with different probability density functions listed in the table below. The purpose of these questions is to practice working with a variety of PDFs and to make computing likelihoods, MLEs, etc. more natural. Note below the *indicator* notation $[x = 0]$ (and $[x = 1]$). The square brackets evaluate to 1 if the argument is true, and 0 otherwise. E.g. if x is 1, the $[x = 0] = 0$ and $[x = 1] = 1$ (here $[x = 0]$ is lazy notation; in Python you would write $x == 0$, for example). We will use the notation a lot, both below and when we learn about classification.

Distribution	$p(x\|\theta)$	Range of x	Range of θ
Bernouilli	$\theta^{[x=1]}(1-\theta)^{[x=0]}$	$x \in \{0, 1\}$	$0 \leq \theta \leq 1$
Beta	$\frac{\Gamma(\theta_1+\theta_0)}{\Gamma(\theta_1)\Gamma(\theta_0)} x^{\theta_1-1}(1-x)^{\theta_0-1}$	$0 \leq x \leq 1$	$\theta_1 > 0, \theta_0 > 0$
Poisson	$\frac{\theta^x}{x!}e^{-\theta}$	$x \in \{0, 1, 2, \ldots\}$	$\theta > 0$
Gamma	$\frac{\theta_1^{\theta_0}}{\Gamma(\theta_0)} x^{\theta_0-1}e^{-\theta_1 x}$	$x \geq 0$	$\theta_1 \geq 0$, $\theta_0 \geq 0$
Gaussian	$\frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{1}{2}\left(\frac{x-\theta_0}{\theta_1}\right)^2}$	$-\infty < x < \infty$	$-\infty < \theta_0 < \infty$, $\theta_1 > 1$

Question 1.1
For each of the probability distributions above, write down their normalizing constants. Remember that $\int p(x|\theta)dx = 1$ for continuous x and $\sum_x p(x|\theta) = 1$ for discrete x.

Question 1.2
You live in Amsterdam and find that it rains quite a lot. You want to estimate the probability that it will rain any given day of the year. Every month for a year you count the number of days with rain, and you get the following (from January to December): 22,19,16,16,14,14,17,18,19,20,21,21 (for a grand total of 217 days with rain).[1] Let r_t be an observation for day t in the year; $r_t = 1$

[1] Source: `http://www.amsterdam.climatemps.com/`.

means there was some rain on day t, $r_t = 0$ means there was no rain. We want to estimate the parameter ρ, the probability of rain on any day of the year. We assume a Bernouilli distribution for the observations $\{r_t\}_{t=1}^{365}$, that is $p(r_t|\rho) = \text{Bernouilli}(r_t|\rho)$. To answer these questions, the number of days of rain per month is not important, only the total for the year is relevant. With this information, answer the following questions:

1. What is the likelihood for a single observation? For the entire set of observations?

2. Write the log-likelihood for the entire set of observations.

3. Solve for the MLE of ρ. Do it in general (with symbols for counts n_0, n_1 for days without and with rain) and for this specific case (plug-in the numbers).

4. Assume a Beta prior for ρ with parameters a and b. What is the MAP for ρ?

5. Write the form of the posterior distribution for ρ? You do not need to solve it analytically.

6. (Optional) Solve for the posterior distribution analytically. Hint: it is a Beta distribution.

Question 1.3
You work in the staffing department of a maternity hospital and part of your job is to determine the staffing requirements during the night shift at your hospital. This might mean the number of doctors and nurses at the hospital and the number of doctors on call (if there are more than the average number of deliveries). Your goal is to determine the distribution over the number of deliveries during the night shift $d_t \in \{0, 1, 2, \ldots\}$ (d for delivery count, t for time, the index of the night). With this you can compute the mean, the probability of more than 5 deliveries, etc. You collect data for two weeks, i.e. $d_1, \ldots, d_{14} = 4, 7, 3, 0, 2, 2, 1, 5, 4, 4, 3, 3, 2, 3$. You assume the observations are explained by a Poisson distribution with parameter λ over the discrete delivery counts. With this information, answer the following questions:

1. What is the likelihood for a single observation? For the entire set of observations?

2. Write the log-likelihood for the entire set of observations.

3. Solve for the MLE of λ. Do it in general and for this specific case (plug-in the numbers).

4. Assume a Gamma prior for λ with parameters a and b. What is the MAP estimate of λ?

5. Write the form of the posterior distribution for λ? (You do not need to solve it analytically)

6. (Optional) Solve for the posterior distribution analytically. Hint: it is a Gamma distribution.

Question 1.4

You have developed a blood test aimed at detecting a disease $d \in \{0, 1\}$ (disease is absent ($d = 0$) or present ($d = 1$)). The test measures the level of a specific indicator of the disease, that is it returns a real valued number relative to some baseline (so the levels can be both negative and positive – anywhere along the real line). Two models of the population are built: one for the patients with the disease, and another for the general population. Measurements tend to have a Gaussian shape, and we therefore model the entire population as a mixture of two Gaussians. That is, $p(l) = p(d = 0)p(l|d = 0) + p(d = 1)p(l|d = 1)$, where $p(d)$ is the prior distribution of patients with and without the disease in the general population and $p(l|d)$ are conditional Gaussian distributions, one for the patients with disease, and one for those without. Note: with this question and the previous two, we are simply applying rules of probability (with some algebra) to get the form of the posterior distribution; however, in this problem we are also classifying (since our target is the discrete label d).

Assume we know $p(d = 0) = \pi_0 = 0.999$ and $p(d = 1) = \pi_1 = 0.001$ from previous experience. We do not know the parameters μ_0, σ_0^2 (the mean and variance of the disease-free population) nor μ_1, σ_1^2 (for the disease population). We measure levels $\{l_n\}_{n=1}^N$ for N people, and we know that $n \in \{D_0\}$ are the indices for the disease free patients and $n \in \{D_1\}$ are the indices for the patients with the disease (i.e. D_0 and D_1 are non-intersecting sets of indices from 1 to N). With this information, answer the following questions:

1. Write down the likelihood of the observations as a product over N level recordings. Hint: use indicator notation (like in the Bernouilli distribution) to distinguish between $d_n = 0$ and $d_n = 1$ in the likelihood.

2. Write down the likelihood as a product over the likelihoods for $\{D_0\}$ and $\{D_1\}$.

3. Compute the log-likelihood.

4. Find the MLE for μ_0 and σ_0^2. Assume we can do the same for μ_1 and σ_1^2

5. We now have our models. To make a prediction, solve for $p(d = 1|l_\star)$, where $l_\star$ is a level recorded for a new patient. Hint: use Bayes theorem.

6. Reduce your solution to have the form of a sigmoid, i.e.

$$p(d = 1|l_\star) = \frac{1}{1 + e^{-a(l_\star)}}.$$

2 Matrix inversion lemma

In computing the posteriors and evidences for Gaussian models, we often encounter complicated forms of the inverse covariance matrices that actually have very simple forms if we apply some matrix manipulation. The matrix inversion lemmas in the Matrix cookbook list a few, in particular, the Woodbury Identity:

$$\left(\mathbf{A} + \mathbf{CBC}^T\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1} + \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{A}^{-1}$$

Question 2.1
When you are working with vector and matrix notation, it can be difficult to know if your solution is correct. A good test is to convert your solution to the scalar case, which is often much simpler to do, and see if the equivalent matrix/vector form matches the scalar form. We will do this for Woodbury by proving the lemma for the scalar case.

1. In the right hand side of the Woodbury identity, replace the matrices with scalars: $\mathbf{A} = a$, $\mathbf{B} = b$, $\mathbf{C} = c$, where a, b, and c are scalars.

2. Prove that the rhs is equal to the lhs for the scalar case.

3 Posterior predictive distributions

Question 3.1
Assume we have 1) observed a single training pair $\{t_1, \boldsymbol{\phi}_1\}$, 2) a Gaussian likelihood (below) and 3) a Gaussian prior over weights $\mathbf{w}$:

$$p(t_1|\boldsymbol{\phi}_1, \mathbf{w}, \beta) = \mathcal{N}\left(t_1|\boldsymbol{\phi}_1^T\mathbf{w}, 1/\beta\right)$$
$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{I}/\alpha\right)$$

Answer the following:

1. Write the posterior distribution $p(\mathbf{w}|t_1, \boldsymbol{\phi}_1, \alpha, \beta)$ as a function of the prior, likelihood, and evidence.

2. Derive the posterior distribution $p(\mathbf{w}|t_1, \boldsymbol{\phi}_1, \alpha, \beta)$.

3. Derive the model evidence $p(t_1|\boldsymbol{\phi}_1, \alpha, \beta)$. Hint: much of the work solving for the posterior can be used to solve for the evidence.

Question 3.2
This question continues from the previous question. Now assume we have observed $N-1$ new data vectors (so that now we have an N-dimensional vector of targets $\mathbf{t}$ and N-by-D dimensional matrix of transformed data $\boldsymbol{\Phi}$) and we have computed the posterior of $\mathbf{w}$:

$$\begin{aligned} p(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \\ \mathbf{m}_N &= \beta \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T\mathbf{t} \\ \mathbf{S}_N &= \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1} \end{aligned}$$

Now we have a test vector $\boldsymbol{\phi}_\star$. We want the posterior predictive distribution for $t_\star$, i.e.

$$p(t_\star|\boldsymbol{\phi}_\star, \boldsymbol{\Phi}, \alpha, \beta) = \int p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) d\mathbf{w} \tag{1}$$

to do this, answer the questions and/or perform the steps below. Note that this is a relatively rare instance in Machine Learning where we can analytically integrate over the model parameters and get an exact form of the posterior (predictive) distribution. Hint: If you get stuck, you can try solving the question using scalar $\phi(\mathbf{x})$ (and w); the steps to the solution will be the same, but will involve scalar operations (division instead of matrix inversion, etc) that are easy. Note: the questions/steps below divide the work required to derive the analytic form of the posterior predictive distribution into manageable chunks; the main question is solve the integral above for $p(t_\star|\boldsymbol{\phi}_\star, \boldsymbol{\Phi}, \alpha, \beta)$. Below, by "Gaussian form", we mean the exponential part of the Gaussian should look like $\exp(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2})$; this is not in "Gaussian form": $\exp(-\frac{1}{2\sigma^2}(x^2 - mx - xm + m^2)$ (here shown for scalar x and m).

1. Compute the joint $p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$. Use correctly labeled C to represent the Gaussian normalizing constants.

2. Rewrite the joint as product of constants and two exponential functions. In one exponential, collect all terms with $\mathbf{w}$; put the remaining terms in the other exponential.

3. Complete the square for $\mathbf{w}$. This will require adding a $\mathbf{m}_{N+1}^T \mathbf{S}_{N+1}^{-1} \mathbf{m}_{N+1}$ to the $\mathbf{w}$ exponential and subtracting from the other (and thus canceling each other).

4. Rewrite the $\mathbf{w}$ exponential so it is in a Gaussian form. Marginalize for $t_\star$ by integrating over $\mathbf{w}$, ensuring the resulting marginal distribution is normalizing (i.e. keep track of all the normalizing constants).

5. What is left is the posterior predictive distribution, but it is not in a Gaussian form; we'll do this now. Collect the squared and linear terms of $t_\star$.

6. The squared term is in the form of $t_\star \beta_\star t_\star$; solve for $\beta_\star$. Hint: use the Woodbury identity. Show the correct solution is $\beta_\star(\boldsymbol{\phi}) = \left(1/\beta + \boldsymbol{\phi}_\star^T \mathbf{S}_N \boldsymbol{\phi}_\star\right)^{-1}$.

7. The linear term is $t_\star \beta_\star y_\star$; solve for $y_\star$.

8. Write the Gaussian form for the posterior predictive distribution using mean and variance functions.

9. (Bonus) Let $\sigma_N^2(\boldsymbol{\phi}) = 1/\beta_\star(\boldsymbol{\phi})$. In the limit $N \to 0$, what is $\sigma_N^2(\boldsymbol{\phi})$? How do you interpret this? How is this related to the *irreducible loss* (similarities and differences)?