STAT 34800 : Modern Methods in Applied Statistics

Spring 2020

•

Midterm

•

Solutions by

Jinhong Du

12243476

There are three questions. You may consult books and existing internet resources, but you may not consult with any people.

The following may be useful in answering question 2.

- We use $Y \sim \text{Poi}(\mu)$ to indicate that Y has a Poisson distribution with both mean and variance equal to μ. Its probability mass function is

$$\Pr(Y = k) = \frac{\mu^k}{k!} e^{-\mu}. \tag{1}$$

- We use $X \sim \text{Gamma}(n, \lambda)$ to indicate that X has a Gamma distribution with mean $\frac{n}{\lambda}$ and density

$$p_X(x) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x}. \tag{2}$$

1

Write a short paragraph (about 4-5 sentences) on each of the following. Each paragraph should begin by describing the key concept or idea. Then provide an example that illustrates the concept. Finally, try to describe where the concept fits into the bigger picture - e.g. why it is important, or what it can be used for.

a) Density estimation

- Density estimation is a procedure that estimates an unknown probability density/mass function of a random variable (or vector) X for a given set of points.
- For example, given data $X_1, \ldots, X_n \in \mathbb{R}$, we divide $\mathbb{R}$ into distinct discrete bins b_j, $j = 1, \ldots, J$. Then we estimate $\Pr(X \in b_j)$ by $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \in b_j}$ and estimate the density in b_j by $\frac{1}{n|b_j|} \sum_{i=1}^n \mathbb{1}_{X_i \in b_j}$ where $|b_j|$ is the length of b_j.
- Density estimation is useful for investing an unknown distribution from data. We can get information such as skewness and multimodality.

b) Overfitting

- Overfitting is a phenomenon that a model performs very well on the training set, while it performs badly on the new samples, i.e., not generalizing well on new samples.
- For example, given a training set with n distinct points in $\mathbb{R}^2$ sample from the curve $y = \sin x$, we can always find a order-n polynomial to perfectly fit these data, however, it is far away from the true underlying curve and it will make many mistakes on new samples.
- Overfitting reminds us that in practice, we should not only consider reducing bias of a model, instead we should try to trade off bias and variance/generality.

c) Cross validation

- Cross validation is a procedure used in model/paramter selection when given a training set. The idea is split training data to some distinct parts by some ways (e.g. leave-one-out, k-fold and etc.), and then each time train a model on some of them and evaulate the model on the rest.
- For example, in leave-one-out validation, we take out one sample in turn at a time to be a *evaulation set*, then train a model on the training set without that sample and evaulate it on the evaulate set. Finally, we look at the average performance in all evaulation sets and determine tha best model/paramters to used.
- It is useful for avoiding overfitting, and in general serve the purpose of improving the generalization ability of the model given limited data.

d) Mixture models

- A mixture model assumes that data come from different distribution $\mathcal{D}_1, \ldots, \mathcal{D}_k$ with probability $\pi_1, \ldots, \pi_k$ respectively, with $\sum_{i=1}^{k} \pi_k = 1$.
- For example, the simplest case is Gaussian Mixture Models with two components, in which we assume that data X comes from $\mathcal{N}(\mu_1, 1)$ and $\mathcal{N}(\mu_2, 1)$ with probability π_1 and π_2 respectively, so that X has density $f(x) = \sum_{i=1}^{2} \pi_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_i)^2}$.
- The mixture models are useful for describing more complex distributions in practice, e.g., multi-modal distributions, by considering the mixture of some simple unimodal distributions.

2

a) Assume that we observe data $X|\lambda \sim \text{Poi}(\lambda)$ and have a prior distribution $\lambda \sim \text{Gamma}(m, l)$. Show that the posterior distribution for λ is also a Gamma distribution, and give expressions for the parameters of this posterior distribution. What is the name for this property, where the posterior distribution has the same parametric form as the prior?

Proof. Since $X|\lambda \sim \text{Poi}(\lambda)$ and $\lambda \sim \text{Gamma}(m, l)$, we have

$$p_{X|\lambda=t}(x) = \frac{t^x}{x!}e^{-t}$$
$$p_\lambda(t) = \frac{l^m}{\Gamma(m)}t^{m-1}e^{-lt}.$$

So

$$p_{\lambda|X=x}(t) \propto p_{X|\lambda=t}(x)p_\lambda(t) \propto t^{m+x-1}e^{-(l+1)t},$$

i.e. $\lambda|X \sim \text{Gamma}(m+x, l+1)$. The property that the posterior distribution has the same parametric form as the prior is called *conjugation*, that is, the Gamma distribution is the conjugate prior for a Poisson mean. □

b) Suppose we observe $X_j|\lambda_j \sim \text{Poi}(\lambda_j)$ for $j = 1, \ldots, n$, and you wish to use shrinkage estimation to improve (point) estimates of λ_j compared with the maximum likelihood estimates.

i) Describe a simple Empirical Bayes (EB) approach to this problem. Give enough detail that each step could be implemented in computer code (but you are not asked to implement them). Give citations for any distributional results that you use.

We assume that $X_j|\lambda_j \overset{iid}{\sim} \text{Poi}(\lambda_j)$ and $\lambda_j|m, l \overset{iid}{\sim} \text{Gamma}(m, l)$.

(1) Compute $(\hat{m}, \hat{l})$, the MLE of (m, l). The log-likelihood for m, l is given by

$$\begin{aligned}
l(m,l) &= \sum_{j=1}^{n} \log p(X_j|m,l) \\
&= \sum_{j=1}^{n} \log \int p(X_j|\lambda_j)p(\lambda_j|m,l)\mathrm{d}\lambda_j \\
&= \sum_{j=1}^{n} \log \int \frac{\lambda_j^{X_j}}{X_j!}e^{-\lambda_j}\frac{l^m}{\Gamma(m)}\lambda_j^{m-1}e^{-l\lambda_j}\mathrm{d}\lambda_j \\
&= \sum_{j=1}^{n} \log \left(\frac{\Gamma(m+X_j)l^m}{X_j!\Gamma(m)} \left(\frac{l}{l+1}\right)^m \left(\frac{1}{l+1}\right)^{X_j} \int \frac{(l+1)^{m+X_j}}{\Gamma(m+X_j)}\lambda_j^{m+X_j-1}e^{-(l+1)\lambda_j}\mathrm{d}\lambda_j \right) \\
&= \sum_{j=1}^{n} [\log(\Gamma(m+X_j)) + m\log(l) - \log(\Gamma(m)) - (m+X_j)\log(l+1)] + \text{constant}
\end{aligned}$$

It turns out that $X_j|m, l \sim \text{Negative-Binomial}\left(m, \frac{1}{l+1}\right)$ and the MLE of (m, l) has no close form. But we can still use numerical algorithms like gradient descent, Newton method and etc, to approximate it.

(2) Obtain point estimate of λ_j from its posterior distribution assuming its prior is $\text{Gamma}(\hat{m}, \hat{l})$.

ii) Following on from i), discuss two different ways you might produce point estimates of λ_j from your EB approach, and how different loss functions for the estimation problem would influence your decision about which to use.

(1) The posterior mode of λ_j given $X = X_j, m = \hat{m}$ and $l = \hat{l}$,

$$\hat{\lambda_j}^{\text{mode}} = \frac{\hat{m} + X_j - 1}{\hat{l} + 1}.$$

since the mode of $\Gamma(\alpha, \beta)$ random variable is $\frac{\alpha-1}{\beta}$ if $\alpha \geq 1$ (https://en.wikipedia.org/wiki/Gamma_distribution).

(2) The posterior mean of λ_j given $X = X_j, m = \hat{m}$ and $l = \hat{l}$,

$$\mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j) = \frac{\hat{m} + X_j}{\hat{l} + 1},$$

since the mean of $\Gamma(\alpha, \beta)$ random variable is $\frac{\alpha}{\beta}$ (https://en.wikipedia.org/wiki/Gamma_distribution).

If we use mean square loss $L(\lambda, \hat{\lambda}) = (\lambda - \hat{\lambda})^2$, then the posterior mean will be preferable, since it minimizes the expected posterior loss,

$$\begin{aligned}
\mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}[L(\lambda_j, \hat{\lambda}_j)] &= \int (\lambda_j - \hat{\lambda_j})^2 p(\lambda_j|X_j) \mathrm{d}\lambda_j \\
&= \mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j^2) - 2\hat{\lambda}_j \mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j) + \hat{\lambda}_j^2 \\
&= \mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j^2) + (\hat{\lambda}_j - \mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j))^2 - [\mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j)]^2 \\
&\geq \mathrm{Var}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j),
\end{aligned}$$

with equality holds if and only if $\hat{\lambda}_j = \mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\lambda_j)$.
Similarly, if we use zero-one loss $L_2(\lambda, \hat{\lambda}) = 1 - \delta(\hat{\lambda} - \lambda)$ where δ is the Dirac delta function (https://en.wikipedia.org/wiki/Dirac_delta_function), then the posterior mode will be preferable, since it minimizes the expected loss,

$$\mathbb{E}_{\lambda_j|X_j}[L_2(\lambda_j, \hat{\lambda}_j)] = \int [1 - \delta(\hat{\lambda} - \lambda)] p(\lambda_j|X_j) \mathrm{d}\lambda_j = 1 - p(\hat{\lambda}_j|X_j) \geq 1 - p(\hat{\lambda}_j^{\text{mode}}|X_j).$$

iii) Consider the following two different data sets A and B:

$$A : (X_1, \ldots, X_n) = (3, 4, 7, 10, 15, 20, 29).$$
$$B : (X_1, \ldots, X_n) = (8, 9, 9, 12, 14, 15, 17).$$

For each data set sketch what you would expect to see if you plotted the EB estimates of $\lambda_1, \ldots, \lambda_n$ (on y axis) vs the maximum likelihood estimates (x axis). Explain the main features of the plots, with particular attention to how they differ, and your reasoning.

The MLE of λ_j is $\hat{\lambda}_j^{MLE} = X_j$ for all j.

- For both A and B, the EB estimator of λ_j increases as X_j inceases. So there will be two increasing curves in the plot.
- As points in A is much more diverse than those in B, we could expect that the curve of A is much steeper than the one of B. This is because that we use X_j's to estimate the MLE of (m, l), and $X_j|m, l$ comes from a negative binomial distribution with parameters $\left(m, \frac{1}{l+1}\right)$.

Solution (cont.)

A will give a smaller estimate of m (the dispersion parameter) and thus a smaller estimate of $p = \frac{1}{l+1}$ since

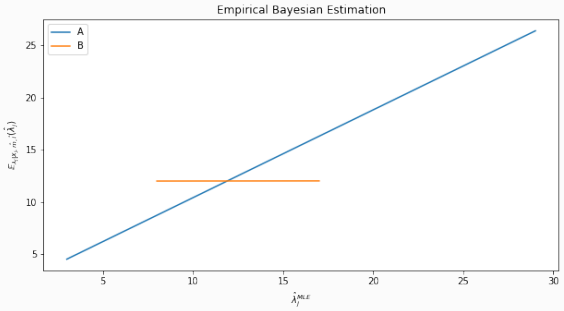
$$\hat{p} = \frac{\sum_{j=1}^{n} X_j}{n\hat{m} + \sum_{j=1}^{n} X_j}, \qquad \hat{l} = \frac{n\hat{m}}{\sum_{j=1}^{n} X_j}$$

by setting the derivative of $l(m, l)$ with respect to p to zero (https://en.wikipedia.org/wiki/Negative_binomial_distribution#Maximum_likelihood_estimation). Then

$$\mathbb{E}_{\lambda_j|X_j}[L(\lambda_j, \hat{\lambda}_j)] = \frac{(\hat{m} + X_j)\sum_{j=1}^{n} X_j}{n\hat{m} + \sum_{j=1}^{n} X_j}.$$

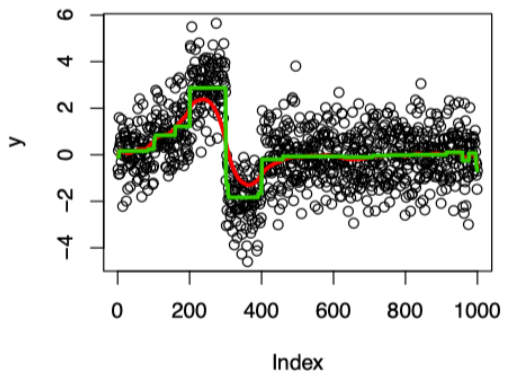
When $\hat{m}$ is very large, this quantity will approach $\frac{1}{n}\sum_{j=1}^{n} X_j$, which may be the case for B, where the points in B are near to each other. So B's EB estimation may look like horizontal.

```
import numpy as np
from scipy.stats import nbinom
from scipy.optimize import minimize
import matplotlib.pyplot as plt
def neg_log_likelihood(pars, X):
    m = np.exp(pars[0])
    l = np.exp(pars[1])
    L = np.sum(nbinom.logpmf(X, m, l/(l+1)))
    return -L
def ebnm_normal(X):
    res = minimize(neg_log_likelihood, np.zeros(2), args=(X), method='L-BFGS-B')
    m, l = np.exp(res.x)
    E_lambda = (m+X)/(l+1)
    return m, l, E_lambda
X1 = np.array([3, 4, 7, 10, 15, 20, 29])
X2 = np.array([8, 9, 9, 12, 14, 15, 17])
plt.figure(figsize=(10,5))
_,_, lambda_hat = ebnm_normal(X1)
plt.plot(X1, lambda_hat, label='A')
_,_, lambda_hat = ebnm_normal(X2)
plt.plot(X2, lambda_hat, label='B')
plt.xlabel('$\hat{\lambda}_j^{MLE}$')
plt.ylabel('$\mathbb{E}_{\lambda_j|X_j,\hat{m},\hat{l}}(\hat{\lambda}_j)$')
plt.legend()
plt.title('Empirical Bayesian Estimation')
plt.show()
```



3

Consider the following graph:



Each point on this graph represents a data point y_i on the y axis, plotted against i on the x axis, for $i = 1, \ldots, 1000$.

The two colored lines show the solutions to the following optimization problems:

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{1000} (y_i - \mu_i)^2 + 0.019 \sum_{k=2}^{1000} |\mu_k - \mu_{k-1}| \tag{3}$$

and

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{1000} (y_i - \mu_i)^2 + 8.08 \sum_{k=2}^{1000} (\mu_k - \mu_{k-1})^2 \tag{4}$$

where in each case $\mu = (\mu_1, \ldots, \mu_{1000})$.

a) Explain what is going on here. Aim your explanation at a statistics student who has not taken this class; explain any technical terms you use.

> Here we are have some data points (x_i, y_i) for $i = 1, \ldots, 1000$. We assume that they are sampled from a curve $y_i = f(x_i) + \epsilon_i$ where ϵ_i is some noise, and the true underlying curve $\mu_i = f(x_i)$ is changing slowly in most regions. Our goal here is to smooth the messy scatter plot and get a smooth line $\hat{\mu}_i$ such as two colored lines above.
> In each optimization problem, the first part $\sum_{i=1}^{1000}(y_i - \mu_i)^2$ describes how well the curve μ_i fits the noisy observation y_i, while the second part $\sum_{k=2}^{1000} |\mu_k - \mu_{k-1}|$ or $\sum_{k=2}^{1000}(\mu_k - \mu_{k-1})^2$ describes the smoothness of the curve μ_i. The number 0.019 or 8.08 is called regularization paramter of the corresponding optimization problem, which trades off the fitness and smoothness of the estimated curve.

b) Which colored line (red and green) corresponds to the fit for each optimization problem? Explain the reasoning behind your answer.

> The green and red lines correspond to the first and second optimization problem respectively.
> The penalty used in the two optimization problems are L_1 penalty and L_2 penalty, respectively. In general, the L_2 penalty will induce a smoother curve than the L_1 penalty, which inspires sparsity so that the estimated curve will look like a staircase curve.