THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

ARIN7101 Statistics in Artificial Intelligence
(2023 Fall)

Assignment 1, due on October 15

All numerical computation MUST be conducted in Python, and attach the Python code.

1. **Question 1 (Bayesian inference, variational inference and sampling)** Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ be i.i.d. samples from the normal distribution $N(\mu, \tau^{-1})$. We specify normal-gamma prior distributions on μ and τ,

$$\mu|\tau, \mu_0, \lambda_0 \sim N(\mu_0, (\lambda_0\tau)^{-1}),$$
$$\tau|a_0, b_0 \sim \text{Gamma}(a_0, b_0),$$
$$f_{\text{Gamma}}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

For a pair of random variables (X, T), if $X|T \sim N(\mu, (\lambda T)^{-1})$ and $T \sim \text{Gamma}(a, b)$, then (X, T) follows a normal-gamma distribution with parameters (μ, λ, a, b). The joint probability density function of (X, T) has the form

$$f(x, t|\mu, \lambda, a, b) = \frac{b^a\sqrt{\lambda}}{\Gamma(a)\sqrt{2\pi}} t^{a-\frac{1}{2}} e^{-bt} \exp\left(-\frac{\lambda t(x-\mu)^2}{2}\right).$$

For the Python programming questions, we set $\mu_0 = 0, \lambda_0 = 10, a_0 = b_0 = 10$ and observations $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ are stored in *Q1y.csv*.

(a) Derive the joint prior (μ, τ) and likelihood function $p(\boldsymbol{y}|\mu, \tau)$. Write down the probability density function of the posterior distribution $(\mu, \tau|\boldsymbol{y})$ (no need to derive the exact distribution)

(b) In fact, for normal distributed data with unknown mean and precision (inverse of variance), the normal-gamma prior is a conjugate prior and the posterior $(\mu, \tau|\boldsymbol{y})$ is also a normal-gamma distribution with parameters

$$\left(\frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n}, \lambda_0 + n, a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{\lambda_0 n(\bar{y} - \mu_0)^2}{2(\lambda_0 + n)}\right),$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is the sample mean.
Derive the full conditional posterior distribution $\mu|\boldsymbol{y}, \tau$ and $\tau|\boldsymbol{y}, \mu$ (need to obtain the exact distribution)

(c) Write down the probability density function of the posterior predictive distribution $y^*|\boldsymbol{y}$. Describe how to approximate $p(y^*|\boldsymbol{y})$ via the simple Monte Carlo approach.

(d) The mode of Normal-Gamma(μ, λ, a, b) is $\left(\mu, \frac{a-\frac{1}{2}}{b}\right)$. Consider the Laplace approximation on the joint posterior $(\mu, \tau|\boldsymbol{y})$:

$$\ln \pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \ln \pi(\boldsymbol{\theta}_{\text{MAP}}|\boldsymbol{y})) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^\intercal \boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}) = \ln \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}),$$
$$\boldsymbol{A} = -\nabla\nabla \ln \pi(\boldsymbol{\theta}|\boldsymbol{y})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MAP}}},$$

where $\boldsymbol{\theta} = (\mu, \tau)^\intercal$.

Derive the approximated posterior distribution $\tilde{\pi}(\mu, \tau|\boldsymbol{y})$. Draw two contour plots of $\pi(\mu, \tau|\boldsymbol{y})$ and $\tilde{\pi}(\mu, \tau|\boldsymbol{y})$ respectively in Python. (Python scipy.stats package does not provide direct functions to calculate the pdf of the normal-gamma distribution. You need to calculate it by yourself)

(e) Assume a mean-field variational inference for the joint posterior $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$. Find the optimal mean-field factor q_μ^* and q_τ^*. Write down the procedures to iteratively update the parameters of q_μ^* and q_τ^* and implement them in Python to obtain the estimated parameters of q_μ^* and q_τ^* (set convergence criterion $\epsilon = 10^{-4}$)
(Hints: $q_j(\theta_j) \propto \exp\{E_{q_{i\neq j}}[\ln P(\mathcal{D}, \boldsymbol{\theta})]\}$)

2. **Question 2 (Regularization)**

Consider the simple linear regression $y_i = \boldsymbol{\beta}^\intercal \boldsymbol{x}_i + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2), i = 1, \ldots, n$, where n is the number of samples, and the residual sum of squares loss,

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\intercal \boldsymbol{x}_i)^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^\intercal(\boldsymbol{y} - \boldsymbol{X\beta}).$$

(a) Under the assumption that $\boldsymbol{X}^\intercal\boldsymbol{X} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, where p is the number of covariates in $\boldsymbol{X}$, derive the closed-form formula for the LASSO regression,

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1,$$

as the function of $\boldsymbol{X}, \boldsymbol{y}, \lambda$ and $(\sigma_1^2, \ldots, \sigma_p^2)$ (do not include $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ in your final results).

(b) The dataset *q2_train.csv* and *q2_test.csv* store age, weight, height, and several body circumference measurements for 252 men. Use the 'brozek' as the response variable (y) and the other variables as predictors ($\boldsymbol{x}$) in the linear regression model.

Normalize the training and test datasets by estimating sample mean and variance from the training dataset. Set $\gamma = 1e-4$ for the learning rate of the proximal gradient method with convergence criteria $\epsilon = 1e-7$.

Plot the estimated coefficients for the ridge regression and LASSO regression, respectively, against $\lambda \in$*np.linspace(0, 350, 1001)*.

(c) Given the LASSO regression results in (b), what's the range of λ if you want to include 4 predictors in the linear regression model? Which four predictors would you choose?

(d) Find the optimal $\lambda \in$ *np.linspace(0, 350, 1001)* which can yield the lowest loss on the test dataset for the LASSO regression. Which predictors are included in the model for the optimal λ?

3. **Question 3 (Multiple Hypothesis Testing)**
The dataset *q3_pvalues.csv* stores two hundred p-values for multiple testings. Consider conducting large-scale hypothesis testing to control the $FWER$ or FDR.

   (a) Conduct Bonferroni's correlation, Holm's procedure, and Benjamini Hochberg's procedure under $\alpha = 0.05$ and $q = 0.05$ then compare the results.

   (b) Overlay the not rejected and rejected p-values, and the corresponding criterion curve in a plot for Bonferroni's correlation, Holm's procedure, and Benjamini Hochberg's procedure **separately**. Denote the not rejected and rejected p-values by different colors. (You can use *log_scale* for demonstration and the plots in our lecture notes are good examples.)