ML HW4 DRY

SNIR HORDAN 205689581, ALEXANDER SHENDER 328626114

1. Consider m i.i.d samples from a normal distribution $x_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown mean and variance.

In tutorial we proved $\hat{\mu}_{MLE} = \overline{X} = \frac{1}{m}\sum\limits_{i=1}^{m} x_i$

Claim. $\hat{\sigma^2}_{MLE} = \frac{1}{m}\sum\limits_{i=1}^{m}\left(x_i - \hat{\mu}_{MLE}\right)^2$

Proof. Calculate log-loss function:

$L(\{x_i\}_{i=1}^m) = -ln(P(\{x_i\}_{i=1}^m)) \underbrace{=}_{x_i\ i.i.d} -ln\left(\prod\limits_{i=1}^{m} P\left(X = x_i\right)\right) = -ln\left(\prod\limits_{i=1}^{m}\left(\sqrt{2\pi}\sigma\right)^{-1}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right)$

$= -ln\left(\left(\sqrt{2\pi}\sigma\right)^{-m}\exp\left\{-\sum\limits_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right) = mln\left(\sqrt{2\pi}\sigma\right) + \sum\limits_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2}$

Find minimum :

$\frac{\partial}{\partial\sigma}mln\left(\sqrt{2\pi}\sigma\right) + \sum\limits_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2} = \frac{m\sqrt{2\pi}}{\sigma\sqrt{2\pi}} - \sum\limits_{i=1}^{m}\frac{(x_i-\mu)^2}{\sigma^3} = 0 \underbrace{\Longrightarrow}_{(1)} \hat{\sigma^2}_{MLE} = \frac{1}{m}\sum\limits_{i=1}^{m}$ $\left(x_i - \hat{\mu}_{MLE}\right)^2$

(1)σ that minimizes log-loss is the maximum likelioood estimator □

Remark: $\hat{\sigma^2}$ is a <u>biased</u> MLE

2.a. $P\left(\boldsymbol{w}|\mu = 0, b\right) \underbrace{=}_{w_i\ i.i.d} \prod\limits_{i=1}^{m}\mathrm{P}(w_i|\mu = 0, b) = (2b)^{-m}\exp\left\{-\frac{1}{b}\sum\limits_{i=1}^{m}|w_i|\right\}$

2.b. $P\left(\boldsymbol{w}|\left\{(x_i, y_i)\right\}_{i=1}^m, \mu = 0, b\right) \underset{Bayes\,Law}{=} P\left(\left\{(x_i, y_i)\right\}_{i=1}^m|\boldsymbol{w}, \mu = 0, b\right) P\left(\boldsymbol{w}|\mu = 0, b\right)\frac{1}{P\left(\left\{(x_i,y_i)\right\}_{i=1}^m, \mu=0, b\right)}$

$= \left[\prod\limits_{i=1}^{m}(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{(y_i-\langle w,x\rangle)^2}{2}\right)\right]\left[(2b)^{-m}\exp\left\{-\frac{1}{b}\sum\limits_{i=1}^{m}|w_i|\right\}\right]\frac{1}{P\left(\left\{(x_i,y_i)\right\}_{i=1}^m, \mu=0, b\right)}$

$\hat{w}_{MAP} := \boldsymbol{argmax}_{\boldsymbol{w}\in\mathbb{R}^d} P\left(\boldsymbol{w}|\left\{(x_i, y_i)\right\}_{i=1}^m, \mu = 0, b\right) = \boldsymbol{argmax}_{\boldsymbol{w}\in\mathbb{R}^d}\ln\left(P\left(\boldsymbol{w}|\left\{(x_i, y_i)\right\}_{i=1}^m, \mu = 0, b\right)\right)$

$= \boldsymbol{argmax}_{\boldsymbol{w}\in\mathbb{R}^d}\ln\left(\left[\prod\limits_{i=1}^{m}(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{(y_i-\langle w,x\rangle)^2}{2}\right)\right]\left[(2b)^{-m}\exp\left\{-\frac{1}{b}\sum\limits_{i=1}^{m}|w_i|\right\}\right]\frac{1}{P\left(\left\{(x_i,y_i)\right\}_{i=1}^m, \mu=0, b\right)}\right)$

$= \boldsymbol{argmax}_{\boldsymbol{w}\in\mathbb{R}^d} - \frac{1}{2}\sum\limits_{i=1}^{m}\left(y_i - \langle w, x\rangle\right)^2 - \frac{1}{b}\sum\limits_{i=1}^{m}|w_i| = \boldsymbol{argmin}_{\boldsymbol{w}\in\mathbb{R}^d}||\boldsymbol{y} - X\boldsymbol{w}||_2^2 +$ $\frac{1}{b}||\boldsymbol{w}||_1^2$

The regularization parameter is therefore: $\lambda = \frac{1}{b}$

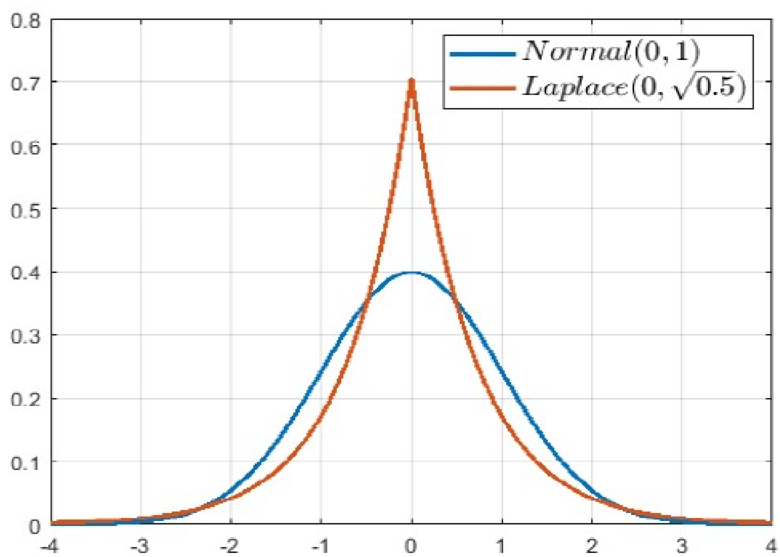
2.c. Ridge regressor corresponds to a MAP estimator under a Gaussian i.i.d prior , $w_i \sim \mathcal{N}\left(0, \frac{1}{\lambda}\right)$

Lasso regressor corresponds to a MAP estimator under a Laplacian i.i.d prior, $w_i \sim Laplace\left(0, \frac{1}{\lambda}\right)$

Let $\lambda = 1 > 0$ be a regularization parameter.

Then $P\left(w_j = 0|\mu = 0, \frac{1}{\lambda}\right) = \frac{\lambda}{2}\exp\left\{-\lambda\left|0\right|\right\} = \frac{\lambda}{2} = \frac{1}{2}$ and $P\left(w_j = 0|\mu = 0, \frac{1}{\lambda}\right) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}$

Then we get the following graph:



Intuitively, there is higher probability of the i-th component of the Lasso regressor being zero then that of the Ridge regressor.

Therefore the Lasso regressor will be more sparse, i.e. more zero components than the Ridge regressor.

3.a

Let $x_1, ..., x_m$ be i.i.d instances drawn from Poisson distribution.

$\hat{\lambda}_{MLE} \coloneqq \underset{\lambda>0}{argmin}\text{-}\ln(P\left(\{x_i\}_{i=1}^m\right))$

$-\ln\left(P\left(\{x_i\}_{i=1}^m\right)\right) \underbrace{=}_{i.i.d} -\ln\left(\prod_{i=1}^m P\left(x_i\right)\right) = -\ln\left(\exp\left\{-m\lambda\right\}\frac{\lambda^{\sum_{i=1}^m x_i}}{\prod_{i=1}^m x_i!}\right) = m\lambda - \sum_{i=1}^m x_i\ln\left(\lambda\right) + \sum_{i=1}^m x_i!$

Find minimum:

$\frac{\partial}{\partial\lambda}m\lambda - \sum_{i=1}^m x_i\ln\left(\lambda\right) + \sum_{i=1}^m x_i! = -m + \frac{1}{\lambda}\sum_{i=1}^m x_i = 0 \implies \hat{\lambda}_{MLE} = \frac{1}{m}\sum_{i=1}^m x_i$

3.b.(i).

The mixture model is a probabilistic model in which the data is samples from K poisson distributions with parameters $\{\lambda_1, ..., \lambda_K\}$ and prior probabilities $\{P(y_1) = c_1, ..., P(y_K) = c_k\}$ of sampling from each distribution.

3.b.(ii).1.

Let $\{x_i\}_{i=1}^m$ be m i.i.d. samples from the mixture model.

Likelihood of incomplete data:

$L(\theta, \{x_i\}_{i=1}^m) = \Sigma_{i=1}^m \ln(P(x_i|\theta)) = \Sigma_{i=1}^m \ln\left(\sum_{j=1}^k P(x_i|y=j,\theta)P(y=j)\right) = \Sigma_{i=1}^m \ln\left(\sum_{j=1}^k \exp\{-\lambda_j\}\frac{x_i^{\lambda_j}}{x_i!}c_j\right)$

where $\theta = \{\lambda_1, ..., \lambda_K, c_1, ..., c_K\}$as defined in 3.b.$(i)$ and $P(x_i|y=j) = \exp\{-\lambda_j\}\frac{x_i^{\lambda_j}}{x_i!}$ by definition of Poisson probability distribution.

Likelihood of complete data:

$\mathrm{L}\big(\theta,\ \{x_i, y_{j_1}\}_{i=1}^m\big) = \sum\limits_{i=1}^m \ln P\left(\{x_i, y_{j_i}\}_{i=1}^m|\theta\right) = \sum\limits_{i=1}^m \ln(P(x_i|y=j)P(y=j)) = \sum\limits_{i=1}^m \ln(P(y_j)) + \ln(P(x_i|y=j))$

$= \sum\limits_{i=1}^m\sum\limits_{j=1}^K \ln(c_j) + \ln\left(\exp\{-\lambda_j\}\frac{x_i^{\lambda_j}}{x_i!}\right)$

where $\theta = \{\lambda_1, ..., \lambda_K, c_1, ..., c_K\}$as defined in 3.b.$(i)$. and $P(x_i|y=j) = \exp\{-\lambda_j\}\frac{x_i^{\lambda_j}}{x_i!}$ by definition of Poisson probability distribution.

3.b.(ii).2.

At Expectation step t+1 the expression Q as defined below is calculated:

$$Q^{(t+1)} = \begin{pmatrix} Q_{11}^{(t+1)} & . & . & . & Q_{1K}^{(t+1)} \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ Q_{m1}^{(t+1)} & . & . & . & Q_{mK|}^{(t+1)} \end{pmatrix} \text{ where } Q_{ij}^{(t+1)} = P\left(y=j \mid x_i, \theta^{(t)}\right) = \frac{P\left(x_i, y=j|\theta^{(t)}\right)}{\sum_j P\left(x_i, y=j|\theta^{(t)}\right)}$$

This defines a new expected log likelihood function over θ.

3.b.(iii) 3.Define $F\left(Q^{(t)}, \theta^{(t)}\right) = \Sigma_i\Sigma_j Q_{ij}^{(t)} \ln\left(P\left(y=j, x_i|\theta^{(t)}\right)\right)$

In lecture we saw, $F\left(Q^{(t)}, \theta^{(t)}\right) - \Sigma_i\Sigma_j Q_{ij}^{(t)} \ln\left(Q_{ij}^{(t)}\right) = \Sigma_i\Sigma_j Q_{ij}^{(t)} \ln\left(P\left(y=j, x_i|\theta^{(t)}\right)\right) - \Sigma_i\Sigma_j Q_{ij}^{(t)} \ln\left(Q_{ij}^{(t)}\right) = l\left(\{x_i\}_{i=1}^m\right)$

Thus, maximizing over $\theta^{(t)}, Q^{(t)}$ improves the log-likelihood of the incomplete data.

3.b.(iii).4.

Define $\theta^{(t+1)}$to be the maximizer of the expected log-likelihood function.

Parameters optimized during maximization step are $\theta^{(t+1)} = \left\{\lambda_1^{(t+1)}, ..., \lambda_K^{(t+1)}\right\}$

$\theta^{(t+1)} = \underset{\theta}{\boldsymbol{argmax}} F\left(Q^{(t)}, \theta\right)$

Derivation for $j \in \{1, ..., K\}$,

$\frac{\partial F\left(Q^{(t)}, \theta^{(t)}\right)}{\partial \lambda_j} = \frac{\partial}{\partial \lambda_j} \Sigma_i \Sigma_j Q_{ij} \ln\left(P\left(x_i | y = j\right) P\left(y = j\right)\right) \underbrace{=}_{(1)} \frac{\partial}{\partial \lambda_j} x_i \ln\left(\lambda_j\right) - \lambda_j + \ln\left(c_j\right) - \ln\left(x_i!\right) = \Sigma_i Q_{ij} \left(\frac{x_i}{\lambda_j} - 1\right) = 0$

$\implies \hat{\lambda_j} = \frac{\Sigma_i x_i Q_{ij}}{\Sigma_i Q_{ij}}$

(1) $\ln\left(\frac{\lambda_j^{x_i}}{x_i!} e^{-\lambda_j} c_j\right)$=$x_iln\left(\lambda_j\right)$-$\lambda_j + \ln\left(c_j\right) - \ln\left(x_i!\right)$