CS771: Practice Set 3

Problem 1

(Learning SVM via Co-ordinate Ascent) Consider the soft-margin linear SVM problem

$$\arg\max_{0 \leq \boldsymbol{\alpha} \leq C} f(\boldsymbol{\alpha})$$

where $f(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}$, $\mathbf{G}$ is an $N \times N$ matrix such that $G_{nm} = y_n y_m \boldsymbol{x}_n^\top \boldsymbol{x}_m$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ are the Lagrange multipliers. Given the optimal $\boldsymbol{\alpha}$, the SVM weight vector is $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$

Your goal is to derive a **co-ordinate ascent** procedure for the vector $\boldsymbol{\alpha}$, such that each iteration updates a uniformly randomly chosen entry α_n of the vector $\boldsymbol{\alpha}$. However, instead of updating $\boldsymbol{\alpha}$ via standard co-ordinate descent as $\alpha_n = \alpha_n + \eta g_n$ where g_n denotes the n-th entry of the gradient vector $\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$, we will update it as $\alpha_n = \alpha_n + \delta_*$ where $\delta_* = \arg\max_\delta f(\boldsymbol{\alpha} + \delta \mathbf{e}_n)$ and $\mathbf{e}_n$ denotes a vector of all zeros except a 1 at entry n.

Essentially, this will give the new α_n that guarantees the maximum increase in f, with all other α_n's fixed at their current value. Derive the expression for δ_* and give a sketch of the overall co-ordinate ascent algorithm.

Note that your expression for δ_* should be such that the constraint $0 \leq \alpha_n \leq C$ is maintained.

Problem 2

(Kernelized Ridge Regression) Starting with the solution of the weight vector of the ridge regression model, i.e., $\boldsymbol{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}$, where $\mathbf{X}$ is the $N \times D$ feature matrix and $\boldsymbol{y}$ is the $N \times 1$ response vector, show that the weight vector can also be written as $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i$ and thus the model can be kernelized at prediction time when given a new test input $\boldsymbol{x}*$.

Problem 3

(MLE for Multivariate Gaussian) Given N samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ drawn i.i.d. from a D-dimensional Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, derive the MLE solution for the mean and covariance matrix of this Gaussian.

Problem 4

(Poisson: Parameter Estimation and Predictive Distribution) Consider N count-valued observations $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ drawn i.i.d. from a Poisson distribution $p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$ where λ is the rate parameter of the Poisson. Assume a gamma prior on λ, i.e., $p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$, where $\alpha > 0$ is the *shape parameter* and $\beta > 0$ is the *rate parameter*, respectively, of the gamma.[1]. Note that, for this parameterization of gamma distribution, the prior's *mode* is $\frac{\alpha-1}{\beta}$ and mean is $\frac{\alpha}{\beta}$.

- Derive the MLE and MAP estimates for λ.
- Derive the posterior distribution for λ.
- Show that the MAP estimate (i.e., mode of the posterior) can be written as weighted combination of the MLE estimate and the prior's mode. Likewise, show that the posterior's *mean* can be written as a weighted combination of the MLE estimate and the prior's *mean*.

[1] There is an alternate parameterization of gamma in terms of shape α and scale θ, for which $p(\lambda) \propto \lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}}$

- Compute the predictive distribution $p(y_*|\boldsymbol{y})$, given the MLE and MAP estimate of λ, and also given the full posterior distribution over λ. In all the three cases, this would be an probability distribution over counts. Is it a Poisson in all the three cases?

Problem 5

(Generative Classification with Gaussian Class-Conditional Distributions) Consider a generative clasification model with K classes. Assume the class-conditional for class k to be Gaussian $\mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k)$ and assume the class-marginal $p(y)$ to be multinoulli$(y|\pi_1, \ldots, \pi_K)$. Assume the training data to be $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$. Show that the MLE solution for the model parameters $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ and $\{\mu_k, \Sigma_k\}_{k=1}^K$ is given by

$$\begin{aligned}
\hat{\pi}_k &= \frac{N_k}{N} \\
\hat{\mu}_k &= \frac{1}{N_k} \sum_{n:y_n=k} \boldsymbol{x}_n \\
\hat{\Sigma}_k &= \frac{1}{N_k} \sum_{n:y_n=k} (\boldsymbol{x}_n - \hat{\mu}_k)(\boldsymbol{x}_n - \hat{\mu}_k)^\top
\end{aligned}$$

Note that this is the same model that we looked at in the class. I however left the derivation as an exercise (which I would like to try on your own now :)). This exercise is also meant to give you practice for parameter estimation for multinoulli and (multivariate) Gaussian distributions.

Note: Although this MLE problem is fairly standard and the results intuitive/worth-remembering, note that:

- For doing MLE for π_k, you will have to use the constraint that $\sum_{k=1}^K \pi_k = 1$ which makes it a *constrained* optimization problem.
- For getting the MLE solution for Σ_k, you should ideally use the constraint that it is positive semi-definite (which will again lead to a constrained optimization problem). However, even if you ignore this constraint, you should be able to get the solution above (so I would suggest ignoring the constraint for this part).

Problem 6

(MAP and Posterior for Multinoulli) You already derived the MLE for the parameters $\boldsymbol{\pi}$ of a multinoulli in Problem 5. In this problem, you will go a step further and do MAP estimation for $\boldsymbol{\pi}$ and also compute the posterior distribution of $\boldsymbol{\pi}$.

Suppose that you are given N samples $y_1, y_2, \ldots, y_N$ drawn i.i.d. from a K-dimensional multinoulli distribution multinoull$(y|\boldsymbol{\pi})$ where $\pi = [\pi_1, \pi_2, \ldots, \pi_K]$. Further, assuming a Dirichlet prior on Dirichlet$(\boldsymbol{\pi}|\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$ are K non-negative hyperparameters of the Dirichlet, derive the MAP estimate as well as the posterior distribution of $\boldsymbol{\pi}$. Note: To get the MAP estimate, you can follow either of these two approaches: (1) Follow the conventional recipe of finding the MAP estimate, i.e., maximize the log-posterior w.r.t. the parameter of interest $\boldsymbol{\pi}$, (2) Compute the posterior of $\boldsymbol{\pi}$ and see what the mode of this distribution is.

Problem 7

(A Circular Definition) Consider a logistic regression model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$, with a zero-mean Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \lambda^{-1}\mathbf{I})$. Note that this loss function for logistic regression assumes $y_n \in \{-1, +1\}$ instead of $\{0, 1\}$. Show that the MAP estimate for $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ where each α_n itself is a function of $\boldsymbol{w}$. Based on the expression of α_n, briefly explain why α_n can be seen as the "importance" of the training example $(\boldsymbol{x}_n, y_n)$ and why the result $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ makes sense for this model.

Problem 8

(Generative seen as Discriminative) Consider a generative classification model for binary classification. Assume the class-marginal distribution to be defined as $p(y = 1) = \pi$ and assume each class-conditional distribution to be defined as a product of D Bernoulli distributions, i.e., $p(\boldsymbol{x}|y = 1) = \prod_{d=1}^{D} p(x_d|y = 1)$ where $p(x_d|y = 1) = \text{Bernoulli}(x_d|\mu_{d,1})$, and $p(\boldsymbol{x}|y = 0) = \prod_{d=1}^{D} p(x_d|y = 0)$ where $p(x_d|y = 0) = \text{Bernoulli}(x_d|\mu_{d,0})$. Note that this makes uses of the naïve Bayes assumption.

Show that this model is equivalent (in its mathematical form) to a probabilistic discriminative classifier. In particular, derive the expression for $p(y = 1|\boldsymbol{x})$, and state what type of decision boundary will this model learn - linear, quadratic, or something else (looking at the expression of $p(y = 1|\boldsymbol{x})$ should reveal that)? Clearly write down the expressions for the parameters of the equivalent probabilistic discrimininative model in terms of the generative model parameters $(\pi, \mu_{d,0}, \mu_{d,1})$. Note that you do not have to estimate the parameters $\pi, \mu_{d,0}, \mu_{d,1}$ (but you may try that for practice if you want).