Intro to Machine Learning (CS771A, Autumn 2018)

Practice Problem Set 3

August 25, 2018

Problem 1

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$(y|\pi_1, \ldots, \pi_K)$. 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 $\sum_{k=1}^K \pi_k = 1$. This is an example of a *constrained* optimization problem that, and one of the ways to solve it is by using *Lagrange multipliers*. If you have seen before how to use Lagrange multiplies in optimization problems, you should be do it. If not, we are going to look at them very soon (Aug 28 lecture), after which you should be able to do it.
- 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 2

Consider the same generative classification model as in Problem 1 and find the MAP estimate for the parameters $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ of the multinoulli class-marginal (you don't need to do it for $\{\mu_k, \Sigma_k\}_{k=1}^K$). For MAP, you need a prior. Assume a K-dimensional Dirichlet prior distribution on $\boldsymbol{\pi}$, given by $p(\boldsymbol{\pi}) = \text{Dirichlet}(\alpha, \ldots, \alpha)$. As I had briefly mentioned in the class, a Dirichlet distribution is a distribution over probability vectors (i.e., vectors that sum to 1), and is sort of a generalization of the Beta distribution (the expression also resembles that of Beta). You may refer to the book/Wikipedia for the expression for the PDF of Dirichlet distribution.

Fun fact: When doing the MAP estimation for $\boldsymbol{\pi}$, you would NOT need to use Lagrange multipliers (unlike in Problem 1 where you have to do MLE)! The Dirichlet prior automatically imposes the sums-to-one constraint.

Note: While you can solve this problem using the standard approach of taking the derivative of the MAP objective, also note that Multinoulli and Dirichlet are also conjugate to each other, so you can even find the full posterior distribution of $\boldsymbol{\pi}$ fairly easily (it will also be Dirichlet due to conjugacy). From the posterior's expression, you can easily get the MAP estimate of $\boldsymbol{\pi}$ (which will be the *mode* of the Dirichlet posterior).

Problem 3

Generative Model for Regression: The probabilistic linear regression model where we directly model a real-valued response as $p(y|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$ can be termed *discriminative* since only y is modeled, not $\boldsymbol{x}$. Suppose we want to construct a *generative* approach to probabilistic linear *regression*. In this approach, instead of directly modeling $p(y|\boldsymbol{x})$, we would like to first model the joint distribution $p(\boldsymbol{x}, y)$ and then use this distribution to obtain our linear regression model $p(y|\boldsymbol{x})$. Assume $\boldsymbol{x}$ to be real-valued vector, i.e., $\boldsymbol{x} \in \mathbb{R}^D$.

Suppose the joint distribution $p(\boldsymbol{x}, y)$ is a $D+1$ dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$. Assume $\mu = \begin{bmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{bmatrix}$ to be the $D+1$ dimensional mean, and $\Sigma = \begin{bmatrix} \Sigma_{xx} & \boldsymbol{v} \\ \boldsymbol{v}^\top & \sigma_y^2 \end{bmatrix}$ to be $(D+1) \times (D+1)$, and assume they have already been estimated (say via MLE/MAP). Also note that in this problem, we are not talking about class-conditional distributions since there are no "classes" so speak of (but we do have assume a joint distribution $p(\boldsymbol{x}, y)$ which is what we need in generative models for supervised learning).

- What will be the conditional distribution $p(y|\boldsymbol{x})$ for this generative regression model? Note that this **will** be a Gaussian (see a result mentioned below). You need to get the expression for the parameters of this Gaussian (basically the mean and variance), for which the result below will be useful.
- Compare the expression of $p(y|\boldsymbol{x})$ obtained for this generative regression model, and compare/constrast with the expression of $p(y|\boldsymbol{x}, \boldsymbol{w}, \beta)$ we have in a probabilistic linear regression model (which is a discriminative model, that doesn't model $\boldsymbol{x}$, unlike the model being considered here, which does model $\boldsymbol{x}$ too). For this comparison, assume Σ_{xx} to be diagonal (this will further simplify the expressions).

Note: This isn't actually a very complicated problem, that will require a lot of algebraic manipulations (especially since I am giving you the result below). The point behind giving this problem is to help you see how regression can also be done using generative models.

Here is the result you will need: If the joint dist. of two sets of variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is Gaussian of the form

$$p\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

then the conditional distribution of one set of variables (say $\boldsymbol{x}_1$) given the other ($\boldsymbol{x}_2$) is also Gaussian, and is given by $p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$ where

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$