CS772A: Probabilistic Machine Learning *Your Name & Roll no:*

Mid-sem Exam (2017-2018, Sem-I)

Instructions:

- This question paper is worth a total of 60 marks.
- Total duration is 3 hours.
- Please write your name and roll number at the top of this sheet as well as on the provided answer copy.
- The true-false questions have to be answered on this question sheet itself, in the space provided. Please also submit the question sheet.
- To answer the other parts of the question paper, please use the provided answer copy.
- For rough work, please use pages at the back of answer copy. Extra sheets can be provided if needed.

Some distributions:

- For $x \in (0,1)$, $\text{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and Γ denotes the gamma function s.t. $\Gamma(x) = (x-1)!$ for a positive integer x.
- For $x \in \{0,1\}$, $\text{Bernoulli}(x|p) = p^x(1-p)^{1-x}$
- For $x \in \mathbb{R}$, Univariate Gaussian: $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
- For $x \in \mathbb{R}^D$, D-dimensional Gaussian: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$.
  Trace-based representation: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\Big\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\Big\}$, $\mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$.
- For $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$, $\text{Dirichlet}(\boldsymbol{\pi}|\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\alpha_1,\ldots,\alpha_K)}\prod_{k=1}^K \pi_k^{\alpha_k-1}$ where $B(\alpha_1,\ldots,\alpha_K) = \frac{\prod_{k=1}^K\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K\alpha_k)}$, and $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^K\alpha_k}$
- For $x_k \in \{0,N\}$ and $\sum_{k=1}^K x_k = N$, $\text{multinomial}(x_1,\ldots,x_K|N,\boldsymbol{\pi}) = \frac{N!}{\boldsymbol{x}_1!\ldots,x_K!}\pi_1^{x_1}\ldots\pi_K^{x_K}$, where $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K\pi_k = 1$. The multinoulli is the same as multinomial with $N=1$.

Some other useful results:

- If $\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon$, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\mathbf{L}^{-1})$ then $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z}+\boldsymbol{b},\mathbf{L}^{-1})$, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu}+\boldsymbol{b},\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top+\mathbf{L}^{-1})$, and $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\boldsymbol{x}-\boldsymbol{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right\},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.
- Marginal and conditional distributions for Gaussians: $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa})$, $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b},\boldsymbol{\Sigma}_{a|b})$ where $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$, $\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)\right\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)$, where symbols have their usual meaning. :)
- $\frac{\partial}{\partial\boldsymbol{\mu}}[\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}] = [\mathbf{A}+\mathbf{A}^\top]\boldsymbol{\mu}$, $\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-\top}$, $\frac{\partial}{\partial\mathbf{A}}\text{trace}[\mathbf{A}\mathbf{B}] = \mathbf{B}^\top$
- For a random variable vector $\boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top + \text{cov}[\boldsymbol{x}]$

Questions begin from the next page

1 True-False Problems (10 questions, 10 marks)

1. [**F**] A linear decision boundary between two Gaussian class-conditional distributions implies that both Gaussians have equal and spherical covariance matrices.(Reason: Only implies equal covariance, not necessarily spherical.)
2. [**F**] Lack of conjugacy in logistic/softmax regression (fixed hyperparameters) makes MAP estimation in these models a non-convex problem. (Reason: Lack of conjugacy only makes Bayesial inference harder but otherwise the loss function for logistic/softmax is convex.)
3. [**T**] Probabilistic PCA can be used for density estimation but non-probabilistic PCA cannot be.
4. [**F**] The EM algorithm for a GMM with identity covariance matrix for each Gaussian and equal mixing proportions will not be sensitive to initialization and will converge to the global optima. (Reason: EM is sensitive to initialization in every case.)
5. [**T**] MLE only gives unregularized solutions and is prone to overfitting.
6. [**F**] In EM, the M step (MLE/MAP) is guaranteed to give a closed form solution of the parameters. (Reason: Only in some cases, the M step may have closed form solution, e.g., when setting derivatives to zero gives a form where you can easily separate the parameters to get a closed form solution.)
7. [**F**] Using the model posterior probability $p(m|\mathcal{D})$ to select the best model (from a set of models $m = 1, \ldots, M$) is equivalent to doing model selection using the marginal likelihood $p(\mathcal{D}|m)$. (Reason: Only if $p(m)$ is uniform, i.e., all models are equally likely *a priori*.)
8. [**F**] We can use importance sampling to compute an intractable expectation $\mathbb{E}_{p(z)}[f(z)]$ only in cases where we can generate samples from the distribution $p(z)$. (Reason: IS works even if we can't sample from $p(z)$ but can evaluate it up to a proportionality constant. How: Using proposal distribution.)
9. [**T**] Assume $x = Az + \epsilon$ where z is a Gaussian random variable and ϵ denotes random noise from an arbitrary distribution. In this case, the distribution of x won't necessarily be a Gaussian. (Reason: The noise must be Gaussian for $p(x|z)$ to be Gaussian.)
10. [**T**] Probabilistic/Bayesian matrix factorization with Gaussian priors on the latent factors can only be called locally conjugate if the matrix entries are also Gaussian when conditioned on the latent factors.

2 Short Answer Problems (10 questions, 20 marks)

Note: Please be concise in your answers (2-4 sentences should suffice to answer each of the questions below).

1. State at least <u>two</u> conditions under which the MAP solution for a probabilistic model would be (at least roughly) equivalent to the MLE solution.

   [Answer: (1) Prior is uniform, (2) Lots of data which "washes out" the effect of the prior.]

2. Suppose you roll a dice 10 times and see outcomes $\{3, 5, 1, 6, 1, 3, 4, 5, 4, 6\}$. Suppose its bias is a 6-dim probability vector $\boldsymbol{\pi}$. Why wouldn't MLE be a good idea to estimate $\boldsymbol{\pi}$? Suggest an alternative (you can answer in plain words!) and briefly justify that choice (rolling the dice more is not an option)?

   [Answer: No outcome with value 2 will make $\pi_2 = 0$ when using MLE. We have too little data, so should use a prior on π and do MAP estimation instead.]

3. What is the difference between a posterior predictive distribution vs a plug-in predictive distribution? State at least <u>two</u> benefits of the former as compared to the latter.

   [Answer: Posterior predictive averages over all values of the parameter (using its posterior distribution) whereas plug-in predictive distribution simply plugs in a single estimate (MLE/MAP) of the parameters. Benefits: The former is more robust to overfitting (since it averages rather than "fitting" a single estimate), (2) Input-dependent variance helps in tasks such as active learning.]

4. For probabilistic linear regression/logistic regression with weight vector $\boldsymbol{w} \in \mathbb{R}^D$, why/when would you want to choose the prior $p(\boldsymbol{w})$ as a Gaussian with diagonal covariance as opposed to a Gaussian with spherical covariance?

   [Answer: Diagonal covariance leads to component-wise regularization on $\boldsymbol{w}$ (different entries can have different amount of regularization), which can be useful in doing feature selection.]

5. Consider probabilistic linear regression and logistic regression, trained using datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. For each of the two models, suppose we have computed the Kullback-Leibler (KL) divergence between the true posterior and its Laplace approximation. Can you say for sure as to which of the two KL divergences will be larger? Justify your answer briefly.

   [Answer: For logistic regression. Laplace approximation approximates the posterior by a Gaussian with mean = MAP estimate and covariance = Hessian (second derivative of the log-likelihood). For linear regression, the KL will be ZERO since Laplace approximation will be the same as exact posterior. For logistic regression, the Laplace approximation isn't equal to the exact posterior, so KL > 0.]

6. Briefly explain how the Bayesian approach to doing inference can inherently work in an online fashion.

   [Answer: Current posterior can be treated as a new prior for the next batch of data. This cycle repeats with every batch of incoming data.]

7. State at least four benefits of using exponential family distributions in probabilistic models.

   [Answer: (1) Have conjugate priors, (2) Finite size sufficient statistics $\Rightarrow$ online inference, (3) MLE is convex (log partition function is convex), (4) Posterior predictive has closed form.]

8. Assuming D-dimensional data, how many *global* parameters do we need to learn in an M component Gaussian mixture model, if each Gaussian has a full covariance matrix? How many *global* parameters do we need to learn in a M component mixture of probabilistic PCA model where the latent factors are K dimensional? Your answer can be in order notation. **Note:** If a parameter is a vector/matrix, the number should be the total size of this vector/matrix (don't count it as a single parameter).

   [Answer: GMM has M (for mixing proportions) + DM (for M means) + $\mathcal{O}(MD^2)$ (for M covariance matrices) parameters. Mixture of PPCA has M (for mixing proportions) + DM (for M means) + MDK (for M factor loading matrices, each of size DK) parameters.]

9. Assuming you have used Laplace approximation to approximate the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$ for a logistic regression model. Why is the posterior predictive $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ still not tractable? Write down its expression and briefly describe a way to approximate it using a sampling method (may use equations).

   [Answer: $p(y_* = 1|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) = \int p(y_* = 1|\boldsymbol{w}, \boldsymbol{x}_*)p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w} = \int \sigma(\boldsymbol{w}^\top \boldsymbol{x}_*)p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w}$. This integral is intractable. We can use sampling to draw L samples $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(L)}$ from the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$ and replace the integral by an empirical average $p(y_* = 1|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) \approx \frac{1}{L}\sum_{\ell=1}^{L} \sigma(\boldsymbol{w}^{(\ell)\top} \boldsymbol{x}_*)$.]

10. Suppose you are given N i.i.d. random variables $x_1, \ldots, x_N$ from $\mathcal{N}(\mu, \sigma^2)$. Consider the random variable $z = \frac{1}{N}\sum_{n=1}^{N} x_n$ and show that it can be written as a linear transformation of a Gaussian random variable. Use this fact and Gaussian linear transformation results to get the distribution of z.

    [Answer: We can write $z = \frac{1}{N}\sum_{n=1}^{N} x_n$ as a linear transformation $z = \boldsymbol{a}^\top \boldsymbol{x}$ where $\boldsymbol{a} = [1/N, \ldots, 1/N]$ and $\boldsymbol{x} = [x_1, \ldots, x_N]$ is an N dimensional Gaussian random variable with distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Using properties of Gaussian lin-trans, mean of $z = \boldsymbol{a}^\top \boldsymbol{\mu} = \mu$. Variance of $z = \boldsymbol{a}^\top \sigma^2 \mathbf{I} \boldsymbol{a} = \sigma^2/N$.]

3 Medium-Length Answer Problems (4 questions, 20 marks)

1. Given i.i.d. training data $(\mathbf{X}, \boldsymbol{y}) = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ from a linear regression model $p(y|\boldsymbol{w}, \boldsymbol{x}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$ with prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})$, show that it is possible to derive the expression for the posterior predictive distribution $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ even **without** using the expression for the posterior distribution of $\boldsymbol{w}$. Show the steps for this derivation and finally write down the expression for $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$.

   [**Answer:** Note that we can also get the posterior predictive without using the posterior of $\boldsymbol{w}$ by first writing down the joint distribution $p(\boldsymbol{y}, y_*|\boldsymbol{x}_*, \mathbf{X})$ and the simply extracting the conditional $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ which is nothing but our posterior predictive. The fact that the marginal likelihood $p(\boldsymbol{y}|\mathbf{X})$ is a Gaussian for linear regression makes it very easy!

   Note that the marginal likelihood $p(\boldsymbol{y}|\mathbf{X}) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = \int \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})d\boldsymbol{w} = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{X}\mathbf{X}^\top + \beta^{-1}\mathbf{I})$ We have seen this result several times (also given on page 1).

   We can use the same result to easily get $p(\boldsymbol{y}, y_*|\boldsymbol{x}_*, \mathbf{X})$. Now the mean $\boldsymbol{\mu}$ will be an $(N+1)$ dimensional zero vector and the covariance matrix $\boldsymbol{\Sigma}$ will be $(N+1) \times (N+1)$ having the same form as $\lambda^{-1}\mathbf{X}\mathbf{X}^\top + \beta^{-1}\mathbf{I}$, but $\mathbf{X}$ replaced by $\tilde{\mathbf{X}}$ where $\tilde{X}$ is the same as $\mathbf{X}$ with $\boldsymbol{x}_*$ appended after the last row.

   We can now apply the conditional from joint formula to get $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ by partitioning $\boldsymbol{\mu}$ (which is simply a zero vector) and $\boldsymbol{\Sigma}$. and using the conditional from joint formula (result on page 1). Even if you have done this much (which is the main crux of the solution) but not written the full expression, you will get full credit.]

2. Consider learning a generative classification model given N i.i.d. training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ with each input $\boldsymbol{x}_n \in \{0, 1\}^D$ and each label $y_n \in \{1, \ldots, K\}$. Assume each class-conditional for $k = 1, \ldots, K$ to be modeled as $p(\boldsymbol{x}|y = k) = \prod_{d=1}^D \text{Bernoulli}(x_d|\theta_{kd})$ where x_d denotes the d-th feature of the input $\boldsymbol{x} \in \{0, 1\}^D$ (I'm skipping subscript n). We would like to do MAP estimation for the parameters θ_{kd}.

   Suggest an appropriate conjugate prior for θ_{kd} (you may assume the same prior for all θ_{kd}'s, $k = 1, \ldots, K$, $d = 1, \ldots, D$) and use this prior to derive the MAP estimate for θ_{kd}.

   [**Answer:** This is the standard naïve Bayes with binary features. You already did the MLE version in HW1 and did a harder version with unknown labels (clustering) using MAP via EM in HW2. So this was supposed to be a very easy question. :) Bernoulli likelihood with a conjugate $\text{Beta}(\theta_{kd}|a, b)$ prior will give you a simple solution for the MAP estimate of $\theta_{kd} = \frac{\sum_{n:y_n=k} x_{nd}+a-1}{\sum_n x_{nd}+a+b-2}$.]

3. Suppose you have rolled a K-sided dice N times. Let us assume each outcome $X_n \in \{1, \ldots, K\}$, $n = 1, \ldots, N$, to be a draw from a multinoulli distribution with parameter vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$. Assume a Dirichlet$(\boldsymbol{\alpha})$ prior on the multinoulli parameter $\boldsymbol{\pi}$ where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ denotes the Dirichlet hyperparameters (assumed known).

   - What will be the posterior distribution of $\boldsymbol{\pi}$, given the observed outcomes $\mathbf{X} = \{X_1, \ldots, X_N\}$? [**Answer:** Due to conjugacy, the posterior $p(\boldsymbol{\pi}|\mathbf{X})$ will be Dirichlet$(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$ where N_k denotes the number of outcomes with $X_n = k$.]
   - What will be the posterior predictive distribution of $(N+1)^{th}$ dice roll, given past outcomes $\mathbf{X}$?

     [**Answer:** Note that we did something like this in the class for the Bernoulli-Beta case. Here $p(X_{N+1} = k|\mathbf{X}) = \int p(X_{N+1}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X})d\boldsymbol{\pi} = \int \pi_k \text{Dirichlet}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)d\boldsymbol{\pi} = \mathbb{E}[\pi_k]$ which is expectation of the multinomial's parameter π_k w.r.t. Dirichlet posterior. This expectation is $\frac{\alpha_k+N_k}{\sum_{k=1}^K(\alpha_k+N_k)} = \frac{\alpha_k+N_k}{\sum_{k=1}^K \alpha_k+N}$ (just in case you didn't remember this result, I gave it on page 1, but your answer will be acceptable even if you only wrote the integral expression with the correct integrand).]

4. Consider the following two step generative model for N binary observations $x_1, \ldots, x_N$: For each $n = 1, \ldots, N$, (1) Draw a binary latent variable $z_n \sim \text{Bernoulli}(\eta)$, (2) If $z_n = 0$, set $x_n = 0$, else generate x_n as $x_n \sim \text{Bernoulli}(\pi)$. For this generative model, derive the expressions for

- Marginal distribution of x_n, i.e., $p(x_n)$ with z_n integrated out.
  [Answer: $p(x_n = 0) = p(x_n = 0|z_n = 0)p(z_n = 0) + p(x_n = 0|z_n = 1)p(z_n = 1) = 1 \times (1 - \eta) + (1 - \pi) \times \eta = 1 - \pi\eta$. And $p(x_n = 1) = \pi\eta$ (Note that the solution makes intuitive sense.)
- Posterior expectation of z_n, assuming η and π are known (e.g., suppose you are doing EM).
  [Answer: Note that $\mathbb{E}[z_n|x_n = 1] = 1$, and therefore we only need to compute $\mathbb{E}[z_n|x_n = 0]$. Note that z_n is binary. Using the basic definition of expectation, $\mathbb{E}[z_n|x_n = 0] = 0 \times p(z_n = 0|x_n = 0) + 1 \times p(z_n = 1|x_n = 0)$. Therefore $\mathbb{E}[z_n|x_n = 0] = p(z_n = 1|x_n = 0) = \frac{p(z_n=1)p(x_n=0|z_n=1)}{p(x_n=0)} = \frac{\eta(1-\pi)}{(1-\eta\pi)}$. (This solution too makes intuitive sense. Can you see why?)

4 Long Answer Problem (1 question, 10 marks)

Assume N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ generated from a D dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We would like to perform MLE to estimate the Gaussian's parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. However, each observation $\boldsymbol{x}_n$ has a part of it as missing (e.g., an image with some "blacked out" pixels). Let us write $\boldsymbol{x}_n = [\boldsymbol{x}_n^{obs}, \boldsymbol{x}_n^{miss}]$ where $\boldsymbol{x}_n^{obs}$ denotes the observed part and $\boldsymbol{x}_n^{obs}$ denotes the missing part. Note that it is not necessary that the same set of features are missing in each observation (e.g., different images may have different parts blacked out).

How can EM be used to solve this problem?

Answer: We can treat the missing observations $\boldsymbol{x}_n^{miss}$ as latent variables in an EM algorithm. The E step would compute the distribution of $\boldsymbol{x}_n^{miss}$ given observed data $\boldsymbol{x}_n^{obs}$ and current estimate of the parameters, and the M step would do MLE on the expected complete data log-likelihood.

Derive the MLE updates for this model using EM. In particular,

1. Write down the expression for the expected complete data log-likelihood (CLL) and simplify it to a form that you can do MLE on.

   Answer: Expected CLL $= \sum_{n=1}^{N} \mathbb{E}[\log p([\boldsymbol{x}_n^{obs}, \boldsymbol{x}_n^{miss}]|\boldsymbol{\mu}, \boldsymbol{\Sigma})]$. Since $p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the expected CLL (ignoring constants) simplifies to $-\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{S}]\right]$, which has the exact same form as the log-likelihood of a multivariate Gaussian, except that we have an expectation of the covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top$ because $\boldsymbol{x}_n$ also has a missing part $\boldsymbol{x}_n^{miss}$.

2. Looking at the expression of expected CLL, what are the required expectations that you will need to compute in the E step? Give the expressions for these quantities.

   Answer: To get the expected CLL, all we need is $\mathbb{E}[\mathbf{S}]$. For the expression of $\mathbb{E}[\mathbf{S}]$, we can write $\mathbb{E}[(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top]$ as $\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top - 2\boldsymbol{\mu}\boldsymbol{x}_n^\top] = \mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top - 2\boldsymbol{\mu}\mathbb{E}[\boldsymbol{x}_n]^\top$. Since $\boldsymbol{x}_n = [\boldsymbol{x}_n^{obs}, \boldsymbol{x}_n^{miss}]$, we can write $\mathbb{E}[\boldsymbol{x}_n] = [\boldsymbol{x}_n^{obs}, \mathbb{E}[\boldsymbol{x}_n^{miss}]]$ and $\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top]$ as

   $$\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top] = \mathbb{E}\begin{bmatrix} \boldsymbol{x}_n^{obs}\boldsymbol{x}_n^{obs\top} & \boldsymbol{x}_n^{obs}\boldsymbol{x}_n^{miss\top} \\ \boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{obs\top} & \boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{miss\top} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_n^{obs}\boldsymbol{x}_n^{obs\top} & \boldsymbol{x}_n^{obs}\mathbb{E}[\boldsymbol{x}_n^{miss}]^\top \\ \mathbb{E}[\boldsymbol{x}_n^{miss}]\boldsymbol{x}_n^{obs\top} & \mathbb{E}[\boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{miss\top}] \end{bmatrix}$$

   So the only expectations we need are $\mathbb{E}[\boldsymbol{x}_n^{miss}]$ and $\mathbb{E}[\boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{miss\top}]$. In the E step, we would compute $p(\boldsymbol{x}_n^{miss}|\boldsymbol{x}_n^{obs}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is simply a Gaussian (using the conditional from joint result given on page 1). $\mathbb{E}[\boldsymbol{x}_n^{miss}]$ will just be the mean of this Gaussian and $\mathbb{E}[\boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{miss\top}] = \mathbb{E}[\boldsymbol{x}_n^{miss}]\mathbb{E}[\boldsymbol{x}_n^{miss}]^\top$ + the covariance matrix of that Gaussian. Even if you have written this much, your answer will be acceptable.

3. Using MLE on expected CLL, derive the expressions for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

   Answer: Note that doing MLE on expected CLL $= -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{S}]\right]$ is almost identical to doing MLE for the Gaussian when $\boldsymbol{x}_n$ is fully observed (which we did in the case in detail), except that in the final solution, we will need to replace all occurrences of quantities such as $\boldsymbol{x}_n, \boldsymbol{x}_n\boldsymbol{x}_n^\top$, etc. by their expectations (computed in Part 2).

Therefore MLE for $\boldsymbol{\mu}$ will be $\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{x}_n]$ and MLE for $\boldsymbol{\Sigma}$ will be $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{S}] = \frac{1}{N}\mathbb{E}[(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top] = \frac{1}{N}\left[\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top - 2\boldsymbol{\mu}\mathbb{E}[\boldsymbol{x}_n]^\top\right] = \frac{1}{N}\mathbb{E}[\boldsymbol{x}_n\boldsymbol{x}_n^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

Note: A common mistake when doing MLE for Gaussian with missing data (or any parameter estimation problem that involves missing data) is to simply take the MLE solution for the non-missing case and only replace $\boldsymbol{x}_n$ by $\mathbb{E}[\boldsymbol{x}_n]$, which is not correct. Note that even if we are directly using the MLE solution of the fully observed data case, we must replace every term that involves missing data by its expectation. Therefore, in this problem, we must do it for both $\boldsymbol{x}_n$ as well as $\boldsymbol{x}_n\boldsymbol{x}_n^\top$ (which is how data appears in the MLE solution). When we properly do EM, our final answer doesn't fall into this trap. :) In the expected CLL, we could clealy see the terms for which we require the expectations, and we found that we must replace both $\boldsymbol{x}_n^{miss}$ and $\boldsymbol{x}_n^{miss}\boldsymbol{x}_n^{miss^\top}$ by their corresponding expectations.

Bonus Problem (5 marks; attempt only after solving the problem above): Repeat the exercise (1) and (2) above but assuming that, instead of a single Gaussian, the data is generated from a mixture of K Gaussians with parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. You need not attempt part (3) but you are welcome to try. :)

Answer: I would encourage you to try this out yourself. You have already seen how to do EM for GMM and how to do MLE for a single Gaussian with missing data (the problem above), so this should not be that hard (and will be a good practice for you).