Monday 20th June 2022, 2pm
(2 hours)

ADVANCED STATISTICAL INFERENCE

Answer all questions.

This examination paper includes three questions and is worth a total of 100 marks.

1. **Probabilistic Reasoning and Modeling** (total of [40] points)

(a) Consider N random variables $\{X_1, \ldots, X_N\}$. Show at least two different ways to write the joint distribution in terms of appropriate marginal and conditional distributions.

[5]

(b) How can you obtain $p(X_2, \ldots, X_N)$ from $p(X_1, \ldots, X_N)$?

[5]

(c) Consider a supervised learning problem with $\mathbf{X}$ and $\mathbf{y}$ representing inputs and labels, respectively. Assume a statistical model with parameters $\mathbf{w}$, that is $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ and a prior over $\mathbf{w}$. Show how to use the product and sum rule to express the marginal likelihood of the model, and discuss what it is useful for.

[5]

(d) Consider the same supervised learning problem as in point **(c)**. What is the meaning of the likelihood and what is the meaning of the prior?

[5]

(e) Consider the same supervised learning problem as in point **(c)**. Is there anything wrong with assuming a prior for the parameters which depends on $\mathbf{X}$, that is $p(\mathbf{w}|\mathbf{X})$, or is this a valid choice for a prior? Explain your answer.

[5]

(f) Consider the same supervised learning problem as in point **(c)**. Imagine using a set of basis functions $\varphi_1(\mathbf{x}), \ldots \varphi_D(\mathbf{x})$ to transform the inputs and consider a Gaussian likelihood $p(\mathbf{y}|\mathbf{w}, \mathbf{X}) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \sigma^2\mathbf{I})$, where Φ is the matrix obtained by applying the D basis functions to the inputs $\mathbf{X}$. Imagine that each basis function $\varphi_j(\mathbf{x})$ is parameterized by parameters ω_j, that is $\varphi_j(\mathbf{x}; \omega_j)$, and that you would like to perform Bayesian inference on all the ω_j. Discuss and derive what functional form you need for $\varphi_j(\mathbf{x}; \omega_j)$ in order to obtain a posterior over all the ω_j in a known form (e.g., Gaussian or other).

[5]

(g) Consider the same supervised learning problem as in the previous point. What do you obtain if you perform a Principal Component Analysis on $\mathbf{X}^\top$ instead of $\mathbf{X}$? Explain what the principal components might mean and what the associated eigenvalues might look like.

[5]

(h) Consider a discrete random variable X with M possible outcomes $x_1, \ldots, x_M$. Consider the entropy of the random variable X assuming a distribution $P(X)$, and use Jensen's inequality to obtain a bound for the entropy $H[X] = -\sum_i P(x_i) \log[P(x_i)]$ as a function of M.
Hint: use the fact that $\log[1/P(x)]$ is convex (positive second derivative).
Bonus question: show that $\log[1/P(x)]$ is indeed convex!

[5]

2. **Kernels** (total of [25] points)

(a) Consider a polynomial kernel $k(x,y) = (\log(x)\log(y) - \log(x^2) - \log(y^2) + 4)$ with $x, y \in \mathbb{R}$. What is the function induced by the kernel? Choose among the following and justify your answer:

1. $\phi(x) = \log(x) + \text{constant}$
2. $\phi(x) = \log(x^2) + \text{constant}$
3. It is infinite dimensional

[5]

(b) Consider an unsupervised learning task with a dataset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in D dimensions. If you apply the K-means algorithm to these data with a transformation of the inputs as $\phi(\mathbf{x}) = \|\mathbf{x} - \mathbf{v}\|^2$ for some $\mathbf{v} \in \mathbb{R}^D$, which one of the following statements is **FALSE**? Justify your answer.

1. K-means returns parabola-shaped clusters in the D-dimensional input space.
2. If $\mathbf{v} = \mathbf{0}$, K-means returns clusters formed by group of points in the D-dimensional space at a similar distance from the origin.
3. The kernel induced by the transformation is $k(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{v}\|^2 \|\mathbf{x}_j - \mathbf{v}\|^2$.

[5]

(c) Explain why some kernels induce an infinite-dimensional mapping. Assuming $x, y \in \mathbb{R}$, explain why the kernel $k(x,y) = \exp(xy)$ induces an infinite-dimensional mapping and why $k(x,y) = \log(xy)$ does not.

[5]

(d) Discuss why Gaussian processes are expensive to use in practice. What are the expensive operations that need to be carried out to optimize Gaussian process parameters and to make predictions?

[5]

(e) Consider a set of N random variables $\mathbf{f} = (f_1, \ldots, f_N)^\top$ associated with N given inputs: $x_1, \ldots, x_N$, and consider modeling $\mathbf{f}$ with a Gaussian process with covariance function:

$$\kappa(x_i, x_j) = \alpha \exp\left(-\beta (x_i - x_j)^2\right)$$

Discuss the effect of α and β on the covariance among the variables $\mathbf{f}$, and what effect this has on the functions that can be drawn from the Gaussian process prior on $\mathbf{f}$.

[5]

3. **Variational Inference** (total of [35] points)

Consider a supervised learning problem with $\mathbf{X}$ and $\mathbf{y}$ representing inputs and labels, respectively. Assume a statistical model with parameters $\mathbf{w}$, that is $p(\mathbf{y}|\mathbf{X},\mathbf{w})$ and a prior over $\mathbf{w}$, and assume making inference over $\mathbf{w}$ using variational inference. Denote by $q(\mathbf{w})$ the approximate posterior.

(a) Explain when variational inference is useful in general, and give one example of a model for which one could use variational inference.

[5]

(b) Write the expression of the evidence lower bound (ELBO) and explain the meaning of each term and symbol in the expression.

[5]

(c) What happens to the ELBO and to the approximate posterior when you apply variational inference with $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu,\Sigma)$ to a model where $p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$ and $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\lambda\mathbf{I})$?

[5]

(d) Assume a prior over $\mathbf{w}$ which factorizes across parameters, that is $p(\mathbf{w}) = \prod_i p(w_i)$ and that $q(\mathbf{w})$ too is factorized $q(\mathbf{w}) = \prod_i q(w_i)$. Assume also that each $p(w_i)$ is parameterized by parameters ψ_i and that we optimize the ELBO with respect to these along with the parameters of the approximate posterior. Explain what happens to the ELBO and to the approximate posterior if we do so.

[5]

(e) Assume that the approximate posterior $q(\mathbf{w})$ is a mixture of Gaussians and the likelihood function $p(\mathbf{y}|\mathbf{X},\mathbf{w})$ is non-Gaussian. Is it easy to apply variational inference in this setting and would it offer a good approximation? Justify your answer.

[5]

(f) Assume that the likelihood function $p(\mathbf{y}|\mathbf{X},\mathbf{w})$ is a mixture of Gaussians and that the approximate posterior $q(\mathbf{w})$ is Gaussian. Is it easy to apply variational inference in this setting and would it offer a good approximation? Justify your answer.

[5]

(g) Consider a classification problem with N observations where $\mathbf{X}$ and $\mathbf{y}$ represent inputs and labels, respectively. Assume a model with a Gaussian process prior over the set of N random variables $\mathbf{f} = (f_1,\ldots,f_N)^\top$, and assume $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p(y_i|f_i)$ with $p(y_i|f_i) = \sigma(f_i)^{y_i}(1-\sigma(f_i))^{1-y_i}$, where $\sigma(a) = \frac{1}{1+\exp(-a)}$. How would you apply variational inference to this setting? Comment on the computational/modeling challenges associated with this approximation.

[5]

END OF QUESTION PAPER