Quiz 0 - STA414/2104 - Solutions

Gaussians

4. If $p(x) = \mathcal{N}(x|\mu, \sigma^2)$, for some $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, can $p(x) < 0$?

   Solution: No. $p(x)$ is a probability density function (PDF), and by definition, a PDF must be non-negative ($p(x) \geq 0$) for all x.

5. If $p(x) = \mathcal{N}(x|\mu, \sigma^2)$, for some $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, can $p(x) > 1$?

   Solution: Yes. Unlike probabilities (mass functions), a probability density can exceed 1. The maximum value of the Gaussian density is $\frac{1}{\sqrt{2\pi\sigma^2}}$ (for $x = \mu$). If the variance is sufficiently small (specifically $\sigma^2 < \frac{1}{2\pi}$ or $\sigma \approx 0.399$), the peak of the density will be greater than 1.

6. If $p(x) = \mathcal{N}(x|\mu, \Sigma)$ with $x \in \mathbb{R}^D$, $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D\times D}$ (a multivariate Gaussian), what is the computational complexity (the asymptotic time cost) of evaluating $p(x)$?

   Solution: $O(D^3)$. Evaluating the density involves inverting the covariance matrix Σ (or computing its Cholesky decomposition). These operations scale cubically with the dimension D.

7. If $p(x) = \mathcal{N}(x|\mu, \Sigma)$ with $x \in \mathbb{R}^D$, $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D\times D}$, what restrictions are there on Σ in order for it to be a valid covariance matrix?

   Solution: Σ must be symmetric and Positive Semi-Definite (PSD).

Derivatives

8. If A is a matrix, what is $\frac{\partial Ax}{\partial x}$?

   Solution: A^T. (This assumes the standard denominator layout where number of rows is equal to the dimension of the input).

9. Given a composition of functions $f(x) = a(b(c(x)))$, we can evaluate its derivative using the chain rule, just multiplying together the Jacobian of each function. What is the fastest order to evaluate this product of Jacobians $J_a \times J_b \times J_c$, when $f(x)$ is a vector-input, scalar-output function?

   Solution: Left-to-Right (Reverse Mode). To illustrate why, consider the dimensions of the Jacobians associated with the intermediate variables. Let the dimensions of the spaces be $x \in \mathbb{R}^N$, $c(x) \in \mathbb{R}^M$, $b(c) \in \mathbb{R}^P$, and $a(b) \in \mathbb{R}^1$. The Jacobian dimensions are:

   - $J_c \in \mathbb{R}^{M\times N}$
   - $J_b \in \mathbb{R}^{P\times M}$
   - $J_a \in \mathbb{R}^{1\times P}$ (Row vector)

   We want to compute the product $J_a J_b J_c$.

   Option 1: Left-to-Right (Reverse Mode)

(a) Compute $v = J_a J_b$. This is a vector-matrix multiplication: $(1 \times P)$ by $(P \times M)$.

(b) **Cost:** $P \times M$ operations. Result has dimensions $1 \times M$.

(c) Compute vJ_c. This is a vector-matrix multiplication: $(1 \times M)$ by $(M \times N)$.

(d) **Cost:** $M \times N$ operations.

(e) **Total Cost:** $PM + MN$.

Option 2: Right-to-Left (Forward Mode)

(a) Compute $M_{bc} = J_b J_c$. This is a matrix-matrix multiplication: $(P \times M)$ by $(M \times N)$.

(b) **Cost:** $P \times M \times N$ operations. Result has dimensions $P \times N$.

(c) Compute $J_a M_{bc}$. This is a vector-matrix multiplication $(1 \times P) \times (P \times N)$.

(d) **Cost:** $P \times N$ operations.

(e) **Total Cost:** $PMN + PN$.

We find that reverse mode is faster as soon as $\frac{M}{M+1} \leq \frac{PN}{P+N}$, and the left side is always < 1. The right side is greater than 1 as soon as $(P-1)(N-1) \geq 1$, which is satisfied as soon as $P, N \geq 2$.

10. How could one form an unbiased estimate of $\nabla_x \int f(x,\theta)p(\theta)d\theta$ given a way to sample from $p(\theta)$ and automatic differentiation?

Solution: By swapping the derivative and the expectation (assuming regularity conditions hold), we have $\nabla_x \mathbb{E}[f(x,\theta)] = \mathbb{E}[\nabla_x f(x,\theta)]$. We can form an unbiased Monte Carlo estimate by sampling $\hat{\theta} \sim p(\theta)$ and computing the gradient of the function at that sample: $\nabla_x f(x, \hat{\theta})$.

Distributions

11. In the exponential family of distributions, $p(x|\theta) = f(x)g(\theta)\exp\{h(x)^T T(\theta)\}$, what must $g(\theta)$ be in order for $p(x|\theta)$ to be a valid probability distribution?

Solution: $g(\theta)$ must be the normalization constant (or the inverse partition function). It ensures the distribution integrates to 1:

$$g(\theta) = \frac{1}{\int f(x)\exp\{h(x)^T T(\theta)\}dx}$$

12. One way to specify a Categorical (discrete) distribution using an unconstrained vector $x \in \mathbb{R}^D$ is with the softmax function: $p(y=c) = \frac{\exp\{x_c\}}{\sum_{c'=1}^{D}\exp\{x_{c'}\}}$. What could go wrong computationally if some elements of x are large?

Solution: Numerical Overflow. If an element x_c is sufficiently large, $\exp(x_c)$ will exceed the maximum representable floating-point number (overflow to infinity), causing the computation to fail or return NaN (erratic behaviour).

13. Regarding the softmax function defined in the previous question: How can you fix this computational issue (when some elements of x are large)?

Solution: Subtract the maximum value $M = \max_i x_i$ from the input vector before exponentiating. Because the softmax is shift-invariant, the result remains the same but overflow is prevented:

$$\frac{\exp\{x_c - M\}}{\sum_{c'=1}^{D}\exp\{x_{c'} - M\}}.$$

The maximal input to the exponential function is (exactly) 0.