STAT215: Assignment 1

Due: January 30, 2020 at 11:59pm PT

Problem 1: *The negative binomial distribution.*

Consider a coin with probability p of coming up heads. The number of coin flips before seeing a 'tails' follows a geometric distribution with pmf

$$\Pr(X = k; p) = p^k (1-p).$$

The number of coin flips before seeing r tails follows a *negative binomial* distribution with parameters r and p.

(a) Derive the probability mass function $\Pr(X = k; r, p)$ of the negative binomial distribution. Explain your reasoning.

(b) The geometric distribution has mean $p/(1-p)$ and variance $p/(1-p)^2$. Compute the mean and variance of the negative binomial distribution. Plot the variance as a function of the mean for fixed p and varying r. How does this compare to the Poisson distribution?

(c) Rewrite the negative binomial pmf in terms of the mean μ and the dispersion parameter r. Show that as $r \to \infty$ with μ fixed, the negative binomial converges to a Poisson distribution with mean μ.

(d) The gamma distribution is a continuous distribution on $(0, \infty)$ with pdf

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\Gamma(\cdot)$ denotes the gamma function, which has the property that $\Gamma(n) = (n-1)!$ for positive integers n. Show that the negative binomial is the marginal distribution over X where $X \sim \text{Poisson}(\mu)$ and $\mu \sim \text{Gamma}(r, (1-p)/p)$, integrating over μ. In other words, show that the negative binomial is equivalent to an infinite mixture of Poissons with gamma mixing measure.

(e) Suppose $X_n \sim \text{NB}(r, p)$ for $n = 1, \ldots, N$ are independent samples of a negative binomial distribution. Write the log likelihood $\mathscr{L}(r, p)$. Solve for the maximum likelihood estimate (in closed form) of $\hat{p}$ for fixed r. Plug this into the log likelihood to obtain the profile likelihood $\mathscr{L}(r, \hat{p}(r))$ as a function of r alone.

Problem 2: *The multivariate normal distribution.*

(a) In class we introduced a multivariate Gaussian distribution via its representation as a linear transformation $x = Az + \mu$ where z is a vector of independent standard normal random variates. Using the change of variables formula, derive the multivariate Gaussian pdf,

$$p(x;\mu,\Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\},$$

where $\mu \in \mathbb{R}^D$ and $\Sigma = AA^\top \in \mathbb{R}^{D\times D}$ is a positive semi-definite covariance matrix.

(b) Let $r = \|z\|_2 = (\sum_{d=1}^D z_d^2)^{1/2}$ where z is a vector of standard normal variates, as above. We will derive its density function.

(i) Start by considering the $D = 2$ dimensional case and note that $p(r)\,\mathrm{d}r$ equals the probability mass assigned by the multivariate normal distribution to the infinitesimal shell at radius r from the origin.

(ii) Generalize your solution to $D > 2$ dimensions, using the fact that the surface area of the D-dimensional ball with radius r is $2r^{D-1}\pi^{D/2}/\Gamma(D/2)$.

(iii) Plot this density for increasing values of dimension D. What does this tell your about the distribution of high dimensional Gaussian vectors?

(iv) Now use another change of variables to derive the pdf of r^2, the sum of squares of the Gaussian variables. The squared 2-norm follows a χ^2 distribution with D degrees of freedom. Show that it is a special case of the gamma distribution introduced in Problem 1.

(c) Rewrite the multivariate Gaussian density in natural exponential family form with parameters J and h. How do its natural parameters relate to its mean parameters μ and Σ? What are the sufficient statistics of this exponential family distribution? What is the log normalizer? Show that the derivatives of the log normalizer yield the expected sufficient statistics.

(d) Consider a directed graphical model on a collection of scalar random variables $(x_1, \ldots, x_D)$. Assume that each variable x_d for $d > 1$ has exactly one parent in the directed graphical model, and let the index of the parent of x_d be denoted by $\mathsf{par}_d \in \{1, \ldots, d-1\}$. The joint distribution is then given by,

$$\begin{aligned} x_1 &\sim \mathcal{N}(0, \beta^{-1}), \\ x_d &\sim \mathcal{N}(x_{\mathsf{par}_d} + b_d; \beta^{-1}) \qquad \text{for } d = 2, \ldots, D. \end{aligned}$$

The parameters of the model are $\beta, \{b_d\}_{d=2}^D$. Show that the joint distribution is a multivariate Gaussian and find a closed form expression the precision matrix, J. How does the precision matrix change in the two-dimensional model where each $x_d \in \mathbb{R}^2$, β^{-1} is replaced by $\beta^{-1}I$, and $b_d \in \mathbb{R}^2$?

Problem 3: *Bayesian linear regression.*

Consider a regression problem with datapoints $(x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$. We begin with a linear model,

$$y_n = w^\mathsf{T} x_n + \epsilon_n; \quad \epsilon_n \sim \mathcal{N}(0, \beta^{-1}),$$

where $w \in \mathbb{R}^D$ is a vector of regression weights and $\beta \in \mathbb{R}_+$ specifies the precision (inverse variance) of the errors ϵ_n.

(a) Assume a standard normal prior $w_i \sim \mathcal{N}(0, \alpha^{-1})$. Compute the marginal likelihood

$$p(\{x_n, y_n\}_{n=1}^N; \alpha, \beta) = \int p(w; \alpha)\, p(\{(x_n, y_n)\}_{n=1}^N \mid w; \beta)\, \mathrm{d}w.$$

(b) Now consider a "spike-and-slab" prior distribution on the entries of w. Let $z \in \{0,1\}^D$ be a binary vector specifying whether the corresponding entries in w are nonzero. That is, if $z_i = 0$ then w_i is deterministically zero; otherwise, $w_i \sim \mathcal{N}(0, \alpha^{-1})$ as above. We can write this as a degenerate Gaussian prior

$$p(w \mid z) = \prod_{i=1}^{D} \mathcal{N}(w_i \mid 0, z_i \alpha^{-1}).$$

Compute the marginal likelihood $p(\{(x_n, y_n)\}_{n=1}^N \mid z, \alpha, \beta)$. How would you find the value of z that maximizes this likelihood?

(c) Suppose that each datapoint has its own precision β_n. Compute the posterior distribution

$$p(w \mid \{(x_n, y_n, \beta_n)\}_{n=1}^N, \alpha).$$

How does the posterior mean compare to the ordinary least squares estimate?

(d) Finally, assume the per-datapoint precisions β_n are not directly observed, but are assumed to be independently sampled from a gamma prior distribution,

$$\beta_n \sim \mathrm{Gamma}(a, b),$$

which has the property that $\mathbb{E}[\beta_n] = a/b$ and $\mathbb{E}[\ln \beta_n] = \psi(a) - \ln b$ where ψ is the digamma function. Then, the errors ϵ_n are marginally distributed according to the Student's t distribution, which has heavier tails than the Gaussian and hence is more robust to outliers.

Compute the conditional distribution $p(\beta_n \mid x_n, y_n, w, a, b)$, and compute the expected log joint

$$\mathcal{L}(w') = \mathbb{E}_{p(\beta_n \mid x_n, y_n, w, a, b)}\left[\log p(\{(x_n, y_n, \beta_n)\}_{n=1}^N, w'; \alpha, a, b)\right].$$

What value of w maximizes the expected log joint probability? Describe an EM procedure to search for,

$$w^* = \arg\max p(w \mid \{(x_n, y_n)\}_{n=1}^N, \alpha, a, b).$$

Problem 4: *Multiclass logistic regression applied to larval zebrafish behavior data.*

Follow the instructions in this Google Colab notebook to implement a multiclass logistic regression model and fit it to larval zebrafish behavior data from a recent paper: `https://colab.research.google.com/drive/1moN5CYNsyxeOSUOmN-QMyqEZwgLSBsjY`. Once you're done, save the notebook in `.ipynb` format, print a copy in `.pdf` format, and submit these files along with the rest of your written assignment.