Exercise 1. *Consider the expected loss for a regression problem under the loss function $l(x, y) = |x - y|$. Show that*

$$\mathbb{E}[l] = \int \int l(y(x), t)\, p(x, t)\, dx\, dt$$

is minimized by the conditional median, i.e. the function $y(x)$ such that the probability mass for $t < y(x)$ is the same as for $t \geq y(x)$.   **2 points**

Exercise 2. *Show that the Dirichlet distribution supported on $\mathbf{p} = (p_1, \ldots, p_K) \subseteq (0,1)^K$ with $\sum_{i=1}^{K} p_i = 1$ and probability density function*

$$p(\mathbf{p}; \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

is the conjugate prior for the categorical distribution on K outcomes:

$$p(X = i \mid \mathbf{p}) = p_i$$

Why are the parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ often called pseudo-counts*?*
Hint: Consider that the likelihood of a data set $D = \{x_n\}_{n=1}^{N}$, where each $x_n \in \{1, \ldots, K\}$, can be written as

$$p(D) = \prod_{k=1}^{K} p_k^{\#\{x_n = k \,:\, x_n \in D\}}$$

2 points

Exercise 3. *Consider a multivariate Gaussian distribution with density*

$$p(\mathbf{x}|\mu, \mathbf{\Sigma}) = (2\pi)^{-\frac{D}{2}} \det |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}$$

Assume that $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$. Then, show that $p(\mathbf{x}_a|\mathbf{x}_b)$ is again a Gaussian distribution and compute its density.

3 points

Exercise 4. *Consider a standard Gaussian distribution in D dimensions:*

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}||x||^2}$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness dr, where $dr \ll 1$, is given by

$$p(r)\, dr \approx S_D r^{D-1} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}r^2}\, dr$$

Here, S_D denotes the surface area of a unit sphere in D dimensions. Show that $p(r)$ has a maximum at $r^* = \sqrt{D-1}$, while the density of $\mathbf{x}$ at this distance to the origin, i.e. $\|\mathbf{x}\| = r$ is smaller than $p(\mathbf{X} = \mathbf{0})$ by a factor of $e^{\frac{D}{2}}$. **3 points**

Exercise 5. *Implement the Bayesian inference for the coin flip example in Python, i.e. write a function which computes the Beta posterior when n samples with k heads have been observed.*

1. *Generate a data set of 100 coin flips which are biased to show 70% heads.*

2. *Show the resulting posterior after $1, 10$ and 100 flips*

3. *Illustrate the role of different priors. Which prior would you choose?*

3 points