# Homework 1
# Bayesian Analysis, Minimax Analysis

CMU 10-716: Advanced Machine Learning (Spring 2020)

OUT: Jan. 30, 2020
DUE: **Feb. 13, 2020, 3:00 PM**.

## Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Bob explained to me what is asked in Question 4.3"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX. Upon submission, label each question using the tempate provided by Gradescope.

# 1 Decision Theory

## 1.1 Minimax Analysis

Let $\mathcal{P}$ be the set of all distributions over $[0, 1]$, and consider the task of estimating the mean of such distributions, with the decision set of possible estimates $\mathcal{A} = \mathbb{R}$. The decision-theoretic loss of interest is the squared loss: $L(p, a) = (\mu_1(p) - a)^2$, where $\mu_1(p)$ is the mean of the distribution $p \in \mathcal{P}$. Suppose that $X_1, \ldots, X_n$ is a sample of size $n$ from some distribution $p \in \mathcal{P}$. Let $X = X_1 + \ldots + X_n$.

1. **(3 points)** Letting $\mu_2(p) = \mathbb{E}_p[X_1^2]$, show that the decision rule $\delta_0(X) = \dfrac{X + \sqrt{n}/2}{n + \sqrt{n}}$ has risk:

$$R(p, \delta_0) = \frac{n(\mu_2(p) - \mu_1(p) + 1/4)}{(n + \sqrt{n})^2}.$$

2. **(3 points)** Show that the worst case risk $\max_{p \in \mathcal{P}} R(p, \delta_0)$ is achieved at and only those distributions $p$ that are concentrated on 0 and 1.
   Hint: the worst case risk is achieved when $(\mu_2(p) - \mu_1(p))$ is maximized.

3. **(6 points)** Prove that $\delta_0$ is minimax.

   (a) First find the minimax risk for Bernoulli parameter estimation.
   Hint: Show that $\delta_0$ is Bayes with respect to a beta prior, and note that a Bayes' estimator with constant risk (independent of p) is a minimax estimator.

   (b) Use this to argue the minimaxity of $\delta_0$ with respect to $\mathcal{P}$.

## 1.2 Bayesian analysis

1. **(4 points)** Suppose that we are interested in estimating some parameter vector $\theta \in \Theta := \mathbb{R}^p$, so that the action or decision set $\mathcal{A} = \mathbb{R}^p$ as well, and suppose that the decision-theoretic loss is the simple quadratic loss:

$$L(\theta, a) = (\theta - a)^T Q (\theta - a)$$

   where $Q$ is a $p \times p$ positive definite matrix. Show that for a prior $\pi$ over $\Theta$, the Bayes estimator of the parameter $\theta$ is

$$\delta^\pi(x) = \mathbb{E}^{\pi(\theta|x)}[\theta].$$

2. **(4 points)** Assume $\theta, x$ and $a$ are real, $\pi(\theta|x)$ is unimodal and symmetric around some point, and $L$ is an increasing function of $|\theta - a|$. Show that the Bayes rule is then the mode of $\pi(\theta|x)$.
   Hint: Find the point of symmetry and then compare the Bayes' risk for any $a$ which is not the median with the Bayes' risk of the median by dividing the real line into suitable parts.

# 2 Understanding Stick Breaking

In this question we will gain some intuition as to why the stick breaking method we saw in class results in a valid sample from the Dirichlet process. For simplicity, we will work with a finite dimensional Dirichlet distribution, but the arguments seen here can be extended to the infinite dimensional Dirichlet process.

Consider the $K$-dimensional Dirichlet distribution, $\mathrm{Dir}(\alpha g)$, with parameters $\alpha > 0$ and probability vector $g$ (i.e. $g_i > 0$ for $i = 1, \ldots, K$ and $\sum_{i=1}^K g_i = 1$). We wish to prove that the following method of constructing $P$ results in a sample from $\mathrm{Dir}(\alpha g)$:

a) Draw $S_1, S_2, \ldots$ independently from Categorical($g$).

b) Draw $V_1, V_2, \ldots$ independently from Beta$(1, \alpha)$

c) Set $P = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1 - V_j)\mathbf{e}_{S_i}$

1. **(5 Points)** Consider a random variable $P \sim \sum_{i=1}^{K} g_i \text{Dir}(\alpha g + \mathbf{e}_i)$. Show that sampling $P$ from this mixture is the same as sampling $P \sim \text{Dir}(\alpha g)$ using the fact that the Dirichlet distribution is conjugate to the categorical distribution. Finally, argue that drawing $S \sim \text{Categorical}(g)$ and $P \sim \text{Dir}(\alpha g + \mathbf{e}_S)$ results in a sample $P \sim \text{Dir}(\alpha g)$.

2. **(5 Points)** An interesting fact about the Dirichlet distribution is that $(z_1, z_2, \ldots, z_K) \sim \text{Dir}(\lambda)$ can be sampled via the following procedure:

   (a) Draw $\gamma_i \sim \text{Gamma}(\lambda_i, 1)$ for $i = 1, \ldots, K$

   (b) Set $z_i = \frac{\gamma_i}{\sum_{j=1}^{K} \gamma_j}$

   Using this fact, show that if $X \sim \text{Dir}(x_1, \ldots, x_K)$, $Y \sim \text{Dir}(y_1, \ldots, y_K)$, $V \sim \text{Beta}(\sum_{i=1}^{K} x_i, \sum_{i=1}^{K} y_i)$, and $Z := VX + (1-V)Y$, then $Z \sim \text{Dir}(x_1 + y_1, \ldots x_K + y_K)$. Assume that we know $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$ are non-negative and constrained such that $\sum_{i=1}^{K} x_i = C_1$ and $\sum_{i=1}^{K} y_i = C_2$ for some constants $C_1, C_2$.

   Hint: It may be useful to know that, given $A \sim \text{Gamma}(\alpha, \theta)$ and $B \sim \text{Gamma}(\beta, \theta)$, $A/(A + B)$ is a Beta distributed with parameters $\alpha$ and $\beta$. Furthermore, where $A_i \sim \text{Gamma}(x_i, \theta)$ and $B_i \sim \text{Gamma}(y_i, \theta)$ for $i = 1, \ldots, K$, $\frac{A_i}{\sum_{j=1}^{K} A_j}$ and $\frac{B_i}{\sum_{j=1}^{K} B_j}$ are independent of $\frac{\sum_{j=1}^{K} A_j}{\sum_{j=1}^{K}(A_j + B_j)}$ for all $i = 1, \ldots, K$. This is due to Basu's theorem which states that any boundedly complete minimal sufficient statistic is independent of any ancillary statistic.

3. **(7 Points)** Use the previous two parts and an inductive argument to show that, for a fixed integer $N \geq 1$, the following procedure:

   (a) Draw $S_1, S_2, \ldots, S_N$ independently from Categorical($g$).

   (b) Draw $V_1, V_2, \ldots, V_N$ independently from Beta$(1, \alpha)$

   (c) Draw $P' \sim \text{Dir}(\alpha g)$

   (d) Set $P = V_1\mathbf{e}_{S_1} + V_2(1 - V_1)\mathbf{e}_{S_2} + \ldots + \left(\prod_{i=1}^{N}(1 - V_i)\right) P'$

   results in $P \sim \text{Dir}(\alpha g)$.

4. **(3 Points)** Why is the above algorithm a bad way to sample from $\text{Dir}(\alpha g)$? Assuming that there was an equivalent algorithm for the infinite dimensional Dirichlet Process (i.e. substituting the infinite-dimensional Dirichlet Process for the finite-dimensional Dirichlet distribution), which of the steps would be infeasible for finite $N$?

# 3 Density Estimation

## 3.1 Total Variation Distance:

Define the Total Variation Distance between distributions $P$ and $Q$ to be:

$$\text{TV}(P, Q) = \sup_{A} \mid P(A) - Q(A) \mid$$

1. **(5 points)** Suppose $P$ and $Q$ have densities $p$ and $q$ respectively, show that:

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int \mid p(x) - q(x) \mid dx.$$

Hint: Consider the sets $S_p = \{x : p(x) > q(x)\}$ and $S_q = \{x : p(x) \leq q(x)\}$.

## 3.2 Leave-One Out Cross-validation Estimator of Risk:

Suppose we have $X_1, .., X_n \sim p$ with $X_i \in (0, 1]$. Let $\widehat{p}_h$ to be the histogram density estimator with bin-width $h = h_n$. Recall that the loss for density estimator $\widehat{p}$ is:

$$L(h) = \int \widehat{p}(x)^2 dx - 2\int \widehat{p}(x)p(x)dx$$

For the second term, we may replace the integral with the average and swap in $\widehat{p}^{(-i)}$ to obtain the leave-one-out cross-validation estimator of risk (as defined on page 12 of the notes):

$$\widehat{R}(h) = \int (\widehat{p}(x))^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{p}^{(-i)}(X_i)$$

In this question, we will compute the leave-one-out cross-validation for the histogram estimator with $d = 1$.

1. **(3 points)** Let $\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \in B_j\}$, show that:

$$\int (\widehat{p}(x))^2 dx = \frac{1}{h^d}\sum_{j=1}^{N} \widehat{\theta}_j^2$$

2. **(7 points)** With that, derive that the cross validation estimator of risk is:

$$\widehat{R}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{N} \widehat{\theta}_j^2$$

Hint: start by relating $\widehat{\theta}_j^{(-i)}$ to $\widehat{\theta}_j$.