The Austalian National University
Research School of Computer Science
Liang Zheng

Semester 2, 2020
Assignment 4 Theory Questions

COMP3670: Introduction to Machine Learning

Note: For the purposes of this assignment, we let lowercase p denote probability density functions (pdf's), and upper case P denote probabilities. If a random variable Z is characterized by a probability density function p, we have that

$$P(a \leq Z \leq b) = \int_a^b p(z)\, dz$$

You should show your derivations, but **you may use a computer algebra system (CAS) to assist with integration or differentiation.**[1].

Question 1 **Bayesian Inference** (40 credits)

Let X be a random variable representing the outcome of a biased coin with possible outcomes $\mathcal{X} = \{0, 1\}$, $x \in \mathcal{X}$. The bias of the coin is itself controlled by a random variable Θ, with outcomes[2] $\theta \in \boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = \{\theta \in \mathbb{R} : 0 \leq x \leq 1\}$$

The two random variables are related by the following conditional probability distribution function of X given Θ.

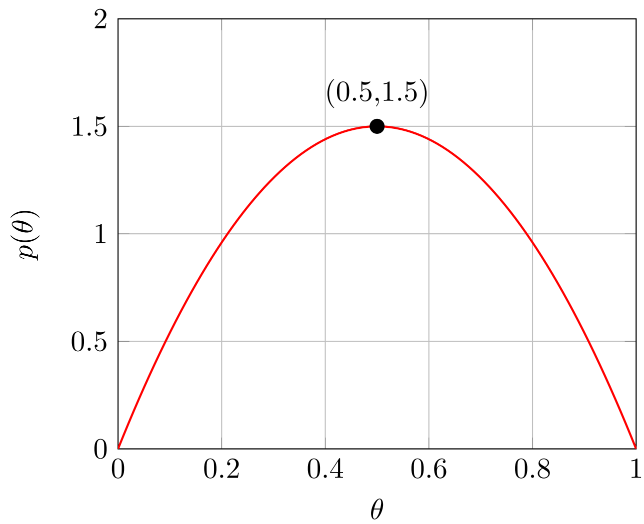
$$p(X = 1 \mid \Theta = \theta) = \theta$$
$$p(X = 0 \mid \Theta = \theta) = 1 - \theta$$

We can use $p(X = 1 \mid \theta)$ as a shorthand for $p(X = 1 \mid \Theta = \theta)$.

We wish to learn what θ is, based on experiments by flipping the coin. Before we flip the coin, we choose as our prior distribution

$$p(\theta) = 6\theta(1 - \theta)$$

which, when plotted, looks like this:



[1] For example, asserting that $\int_0^1 x^2 \left(x^3 + 2x\right)\, dx = 2/3$ with no working out is adequate, as you could just plug the integral into Wolfram Alpha using the command `Integrate[x^2(x^3 + 2x),{x,0,1}]`

[2] For example, a value of $\theta = 1$ represents a coin with 1 on both sides. A value of $\theta = 0$ represnts a coin with 0 on both sides, and $\theta = 1/2$ represents a fair, unbaised coin.

a) (3 credits) Verify that $p(\theta) = 6\theta(1-\theta)$ is a valid probability distribution on $[0,1]$ (i.e that it is always non-negative and that it is normalised.)

We flip the coin a number of times.[3] After each coin flip, we update the probability distribution for θ to reflect our new belief of the distribution on θ, based on evidence.

Suppose we flip the coin twice, and obtain the sequence of coin flips [4] $x_{1:2} = 00$. For each subsequence $x_1, x_{1:2}$ (and for the case before any coins are flipped), compute the:

b) (15 credits) probability distribution functions

c) (3 credits) expectation values μ

d) (3 credits) variances σ^2

e) (5 credits) The *maximum a posteriori* estimation θ_{MAP}.

Present your results in a table like as shown below.

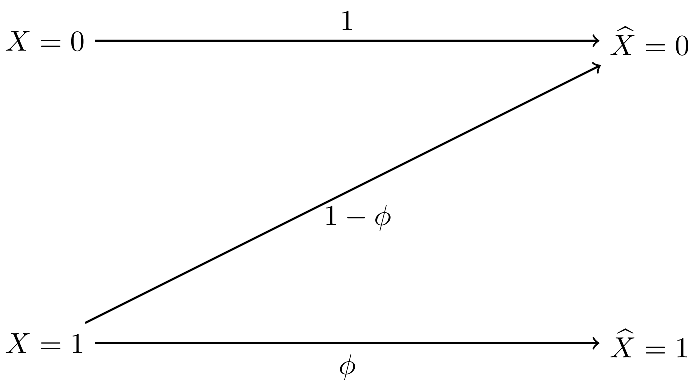
Posterior	**PDF**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\theta_{MAP}}$
$p(\theta)$	$6\theta(1-\theta)$	?	?	?
$p(\theta \mid x_1 = 0)$	?	?	?	?
$p(\theta \mid x_{1:2} = 00)$	?	?	?	?

f) (5 credits) Plot each of the probability distributions $p(\theta), p(\theta|x_1 = 0), p(\theta|x_{1:2} = 00)$ over the interval $0 \leq \theta \leq 1$ on the same graph to compare them.

g) (6 credits) What behaviour would you expect of the posterior distribution $p(\theta|x_{1:n})$ if we updated on a very long sequence of alternating coin flips $x_{1:n} = 10101010\ldots$?
What would you expect $\mu, \sigma^2, \theta_{MAP}$ to look like for large n?
Sketch/draw an estimate of what $p(\theta|x_{1:n})$ would approximately look like against the other distributions.

Question 2 **Bayesian Inference on Imperfect Information** (50 credits)

We have a Bayesian agent running on a computer, trying to learn information about what the parameter θ could be in the coin flip problem, based on observations through a noisy camera. The noisy camera takes a photo of each coin flip and reports back if the result was a 0 or a 1. Unfortunately, the side of the coin with a "1" on it is very shiny, and the reflected light causes the camera to sometimes report back the wrong result.[5] The probability that the camera correctly reads a one is parameterised by $\phi \in [0,1]$. The camera always correctly identifies zeros. Letting X denote the true outcome of the coin, and $\widehat{X}$ denoting what the camera reported back, we can draw the relationship between X and $\widehat{X}$ as shown.



[3] The coin flips are independent and identically distributed (i.i.d).
[4] We write $x_{1:n}$ as shorthand for the sequence $x_1 x_2 \ldots x_n$.
[5] The errors made by the camera are i.i.d, in that past camera outputs do not affect future camera outputs.

So, we have

$$\begin{aligned}
p(\widehat{X} = 0 \mid \phi, X = 0) &= 1 \\
p(\widehat{X} = 0 \mid \phi, X = 1) &= 1 - \phi \\
p(\widehat{X} = 1 \mid \phi, X = 1) &= \phi \\
p(\widehat{X} = 1 \mid \phi, X = 0) &= 0
\end{aligned}$$

We would now like to investigate what posterior distributions are obtained, as a function of the parameter ϕ. Let $\widehat{x}_{1:n}$ be a sequence of coin flips as observed by the camera.

a) (5 credits) Briefly comment about how the camera behaves for $\phi = 0, \phi = 0.5, \phi = 1$. How you expect this would change how the agent updates it's prior to a posterior on θ, given an observation of $\widehat{X}$. (No equations required.)

b) (10 credits) Compute $p(\widehat{X} = x|\theta, \phi)$ for all $x \in \{0, 1\}$.

c) (15 credits) The coin is flipped, and the camera reports seeing a zero. (i.e. that $\hat{x}_1 = 0$.)
Given the same choice of prior $p(\theta|\phi) = 6\theta(1 - \theta)$ as before, compute the posterior $p(\theta|\widehat{x}_1 = 0, \phi)$. What term (from Question 1) does $p(\theta|\widehat{x}_1 = 0, \phi)$ simplify to when $\phi = 1$? When $\phi = 0$? Explain your observations.

d) (10 credits) The experiment is reset.
The coin is flipped, and the camera reports seeing a one. (i.e. that $\hat{x}_1 = 1$.)
Given the same choice of prior $p(\theta|\phi) = 6\theta(1 - \theta)$ as before, compute the posterior $p(\theta|\widehat{x}_1 = 1, \phi)$. Comment on how the result depends on ϕ. Does the result make sense?

e) (10 credits) Plot $p(\theta|\widehat{x}_1 = 0, \phi)$ as a function of θ, for all $\phi \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ on the same graph to compare them. Comment on how the shape of the distribution changes with ϕ. Explain your observations.

Question 3 **Relating Random Variables** (10 credits)

Let X be a random variable, on $[0, 1]$, with probability density function

$$p(x) = 2 - 2x$$

Let Y be a random variable on $[1, 2]$, such that $Y = X^2 + 1$. Find the probability density function for Y.