**The Austalian National University**
**School of Computing**
Jo Ciucă & Thang Bui

Semester 2, 2023
Tutorial 8

# COMP3670/6670: Introduction to Machine Learning

## Question 1 Bayesian linear regression, one parameter, one training point

Consider a linear regression problem with a single training point (x, y) where $x, y$ are scalars, and an one-dimensional parameter $\theta$ (with no bias/intercept):

$$\text{prior:} \quad p(\theta) = \mathcal{N}(\theta; m_o, v_o) \tag{1}$$

$$\text{likelihood:} \quad p(y|\theta, x) = \mathcal{N}(y; \theta x, v) \tag{2}$$

a) Find the marginal likelihood $p(y|x) = \int \mathrm{d}\theta p(y|\theta, x)p(\theta)$

b) Find the posterior distribution $p(\theta|y, x) = \frac{p(y|\theta,x)p(\theta)}{p(y|x)}$

c) Assuming $x \neq 0$, what happens when $v_o$ is very large?

d) What happens when $x = 0$?

**Solution.**

a) We first write down the prior and conditional densities:

$$p(\theta) = \mathcal{N}(\theta; m_o, v_o) = \frac{1}{\sqrt{2\pi v_o}} \exp\left(-\frac{1}{2}\frac{(\theta - m_o)^2}{v_o}\right), \tag{3}$$

$$p(y|\theta, x) = \mathcal{N}(y; \theta x, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{1}{2}\frac{(y - x\theta)^2}{v}\right). \tag{4}$$

We then write down the joint, which is the product of these densities

$$p(y, \theta|x) = p(\theta)p(y|\theta, x) \tag{5}$$

$$= \frac{1}{\sqrt{2\pi v_o}} \exp\left(-\frac{1}{2}\frac{(\theta - m_o)^2}{v_o}\right) \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{1}{2}\frac{(y - x\theta)^2}{v}\right). \tag{6}$$

Noting that we need to integrate out $\theta$, so we will group the terms that have $\theta$ together,

$$p(y, \theta|x) = \frac{1}{\sqrt{2\pi v_o}} \exp\left(-\frac{\theta^2}{2v_o} + \frac{\theta m_o}{v_o} - \frac{m_o^2}{2v_o}\right) \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{y^2}{2v} + \frac{\theta xy}{v} - \frac{\theta^2 x^2}{2v}\right) \tag{7}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi v_o}} \exp\left(-\frac{m_o^2}{2v_o}\right) \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{y^2}{2v}\right)}_{C} \exp\left(\underbrace{-\frac{\theta^2}{2v_o} + \frac{\theta m_o}{v_o} + \frac{\theta xy}{v} - \frac{\theta^2 x^2}{2v}}_{A}\right) \tag{8}$$

Consider the quadratic term involving $\theta$ inside the exponential:

$$A = -\theta^2\left(\frac{1}{2v_o} + \frac{x^2}{2v}\right) + \theta\left(\frac{1}{v_o} + \frac{x^2}{v}\right)\frac{m_o v + xyv_o}{v + x^2 v_o} \tag{9}$$

$$= -\frac{1}{2}\left(\frac{1}{v_o} + \frac{x^2}{v}\right)\left[\theta - \frac{m_o v + xyv_o}{v + x^2 v_o}\right]^2 + \left(\frac{m_o v + xyv_o}{v + x^2 v_o}\right)^2 \left(\frac{1}{2v_o} + \frac{x^2}{2v}\right) \tag{10}$$

$$= -\frac{1}{2}\left(\frac{1}{v_o} + \frac{x^2}{v}\right)\left[\theta - \frac{m_o v + xyv_o}{v + x^2 v_o}\right]^2 + \frac{(m_o v + xyv_o)^2}{2(v + x^2 v_o)vv_o}. \tag{11}$$

This leads to,

$$p(y,\theta|x) = C\exp(A) \tag{12}$$

$$= C\exp\left(\frac{(m_o v + xyv_o)^2}{2(v+x^2v_o)vv_o}\right)\sqrt{2\pi\frac{v_o v}{x^2v_o+v}}\mathcal{N}\left(\theta;\frac{m_o v+xyv_o}{v+x^2v_o},\frac{vv_o}{v+x^2v_o}\right) \tag{13}$$

As the Gaussian density integrates to 1 and the term in front does not depends on $\theta$,

$$p(y|x) = \int \mathrm{d}\theta p(y,\theta|x) \tag{14}$$

$$= C\exp\left(\frac{(m_o v + xyv_o)^2}{2(v+x^2v_o)vv_o}\right)\sqrt{2\pi\frac{v_o v}{x^2v_o+v}} \tag{15}$$

$$= \frac{1}{\sqrt{2\pi(x^2v_o+v)}}\exp\left(-\frac{1}{2}\frac{(y-m_o x)^2}{x^2v_o+v}\right) \tag{16}$$

$$= \mathcal{N}(y;m_o x, x^2v_o+v). \tag{17}$$

b) From equations 13 and 17,

$$p(\theta|y,x) = \mathcal{N}\left(\theta;\frac{m_o v+xyv_o}{v+x^2v_o},\frac{vv_o}{v+x^2v_o}\right). \tag{18}$$

c) Assuming $x \neq 0$, when $v_o$ is very large, the posterior becomes:

$$p(\theta|y,x) = \mathcal{N}\left(\theta;\frac{y}{x},\frac{v}{x^2}\right), \tag{19}$$

that is the posterior mean is the maximum likelihood estimate $y/x$.

d) When $x = 0$, the posterior becomes,

$$p(\theta|y,x) = \mathcal{N}\theta;m_o,v_o), \tag{20}$$

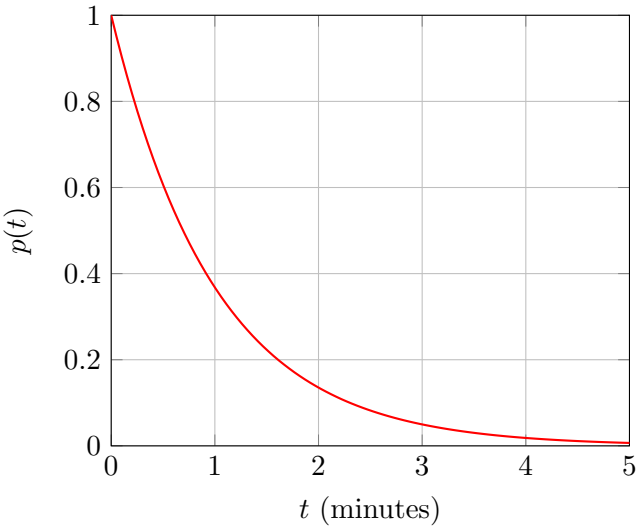
which is the prior. In this case, the observation at $x = 0$ does not bring any new information about which line (without bias) approximately goes through $(x, y)$.

**Question 2** **Bomb Defusal**

You are a bomb defusal specialist, and you've come across a bomb that has just armed itself. You know from experience with these kinds of bomb, that the time till it explodes is controlled by a random variable $T$, sampled from the interval $[0, +\infty)$, according to the prior pdf function

$$p(t) = e^{-t} \text{ minutes}$$

which when plotted, looks like this



a) How likely is it for the bomb to take between 1 and 2 minutes to explode?

**Solution.** We compute $P(1 \leq t \leq 2)$,

$$P(1 \leq t \leq 2) = \int_1^2 e^{-t}\, dt = \frac{1}{e} - \frac{1}{e^2} \approx 23.3\%$$

b) How long on average until the bomb explodes?

**Solution.** The average time to detention is given by the expected value of the pdf, which is (using a computer algebra system)

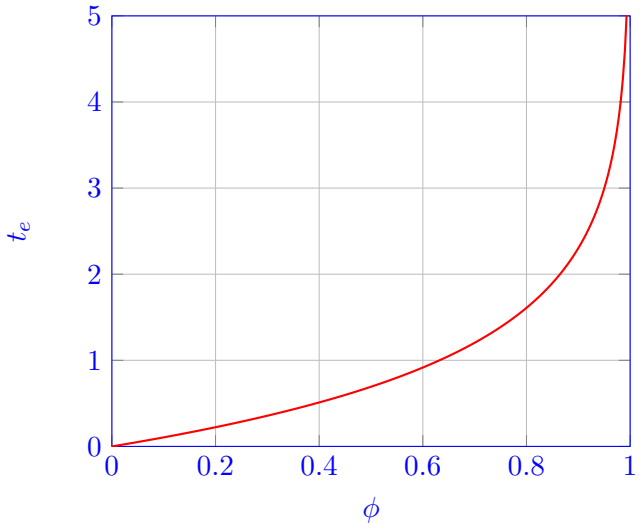
$$\mathbb{E}_T[t] = \int_0^\infty tp(t)\, dt = \int_0^\infty te^{-t} dt = 1 \text{ minute}$$

c) For any $\phi \in [0, 1)$, how much time $t_\phi$ would have to pass such that the probability of the bomb having exploded by then is $\phi$?

**Solution.** We need to compute $p(t \leq t_e) = \phi$, and solve for $t_e$.

$$\begin{aligned}
p(t \leq t_e) &= \phi \\
\int_0^{t_e} e^{-t} dt &= \phi \\
-\, e^{-t}\big|_0^{t_e} &= \phi \\
1 - e^{-t_e} &= \phi \\
t_e &= \log \frac{1}{1-\phi}
\end{aligned}$$

When plotted, it looks like this




d) Suppose you waited a minute, and the bomb has not yet exploded. What is the posterior distribution $p(t \mid t \geq 1)$ of the bomb's detonation time? Plot $p(t \mid t \geq 1)$ against the prior $p(t) = e^{-t}$.

**Solution.** Apply Bayes' rule.

$$p(t \mid t \geq 1) = \frac{p(t \geq 1 \mid t)p(t)}{p(t \geq 1)} = \frac{p(t \geq 1 \mid t)p(t)}{\int_0^\infty p(t \geq 1, t)\, dt} = \frac{p(t \geq 1 \mid t)p(t)}{\int_0^\infty p(t \geq 1 \mid t)p(t) dt}$$

Now, what is $p(t \geq 1 \mid t)$? It's the probability distribution of $t \geq 1$ given the value of $t$. Well, if $t \geq 1$, then the probability of $t \geq 1$ being true is 100%, and if $t < 1$, then the probability of $t \geq 1$ being true is 0%. Hence,

$$p(t \geq 1 \mid 1) = \begin{cases} 1 & t \geq 1 \\ 0 & t < 1 \end{cases}$$

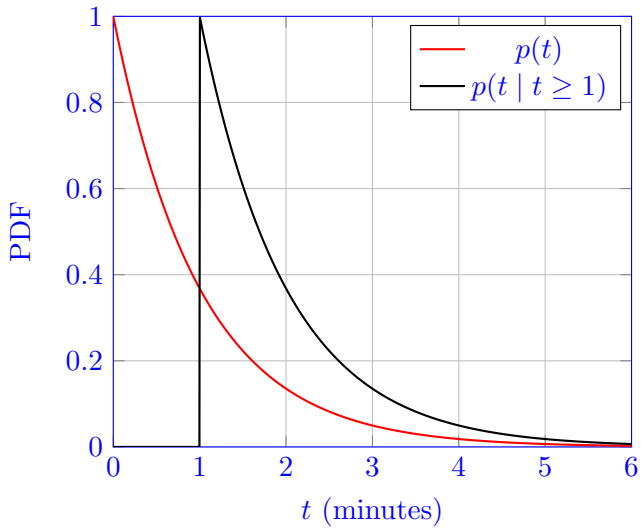
Hence,

$$\begin{aligned} &= \frac{p(t \geq 1 \mid t)p(t)}{\int_0^\infty p(t \geq 1 \mid t)p(t)dt} \\ &= \frac{p(t \geq 1 \mid t)p(t)}{\int_0^1 p(t \geq 1 \mid t)p(t)dt + \int_1^\infty p(t \geq 1 \mid t)p(t)dt} \\ &= \frac{p(t \geq 1 \mid t)p(t)}{\int_0^1 0p(t)dt + \int_1^\infty 1p(t)dt} \\ &= \frac{p(t \geq 1 \mid t)e^{-t}}{\int_1^\infty e^{-t}dt} \end{aligned}$$

Note that $\int_1^\infty e^{-t}dt = e^{-1}$,

$$\begin{aligned} &= ep(t \geq 1 \mid t)e^{-t} \\ &= p(t \geq 1 \mid t)e^{-t+1} \\ &= \begin{cases} e^{-t+1} & t \geq 1 \\ 0 & t < 1 \end{cases} \end{aligned}$$

Plotting, we obtain




e) You're a fast defuser, but an even faster runner. The bomb is placing 5 other people in mortal danger. It would take you 15 seconds to move out of the blast radius of the bomb, and 90 seconds to complete a defusal of the bomb. What action maximizes the expected number of lives saved? Attempting a defuse, or running away?

**Solution.** Running away means the total number of lives saved is only that of the defuser, multiplied by the probability that they get away in time. (that is, the probability that the bomb takes longer than 0.25 minutes to explode. So, on expectation, the number of lives saved is

$$1 \cdot p(t \geq 0.25) = \int_{0.25}^\infty e^{-t}dt \approx 0.779$$

Attempting to defuse means the total number of lives saved is 6 (the 5, plus the defuser), times the probability of a successful defuse (that is, the probability of the bomb taking longer than 1.5 minutes to explode.) So on expectation, the number of lives saved is

$$6 \cdot p(t \geq 1.5) = 6 \int_{1.5}^{\infty} e^{-t} dt \approx 1.33$$

So defusing appears to be on expectation more lives saved than running away.

f) You discover a bomb just as it arms itself. This bomb is equipped with a display, reading out the time left till detonation in minutes in seconds (e.g the display says `1: 30` for 90 seconds). Unfortunately, part of the display is damaged, and you can only read the first digit (the minutes digit), which is a 1. What is the posterior $p(t \mid \text{First digit is 1})$ based on this evidence? Plot this against $p(t)$.

**Solution.** Apply Bayes' rule. Note that if the first digit of the bomb is a 1, the time left $t$ must satisfy $1 \leq t < 2$

$$p(t \mid \text{ First digit is 1}) = \frac{p(\text{ First digit is 1} \mid t)p(t)}{p(\text{ First digit is 1})} = \frac{p(1 \leq t < 2 \mid t)e^{-t}}{\int_0^{\infty} p(1 \leq t < 2 \mid t)e^{-t} dt}$$

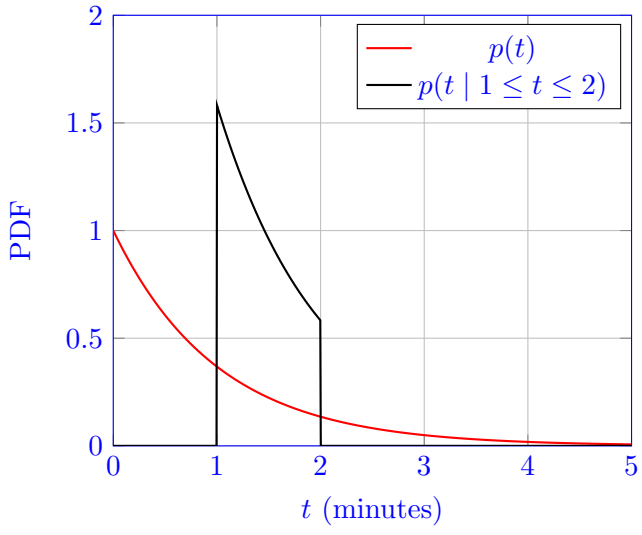
As before, note that $p(t \mid \text{ First digit is 1})$ is 1 if $1 \leq t \leq 2$, and 0 otherwise. Hence,

$$\begin{aligned}
&\frac{p(\text{ First digit is 1} \mid t)p(t)}{p(\text{ First digit is 1})} \\
&= \frac{p(1 \leq t < 2 \mid t)e^{-t}}{\int_0^{\infty} p(1 \leq t < 2 \mid t)e^{-t} dt} \\
&= \frac{p(1 \leq t < 2 \mid t)e^{-t}}{\int_1^2 e^{-t} dt}
\end{aligned}$$

Note that $\int_1^2 e^{-t} dt = \frac{e-1}{e^2}$,

$$\begin{aligned}
&= \frac{e^2}{e-1} p(1 \leq t < 2 \mid t)e^{-t} \\
&= \begin{cases} \frac{e^2}{e-1} e^{-t} & 1 \leq t \leq 2 \\ 0 & \text{else} \end{cases}
\end{aligned}$$

Plotting the results,



g) Suppose two bombs[1] (with the same distribution $p(t) = e^{-t}$ as before) are armed simultaneously. Let $T_1, T_2$ denote random variables for the detonation time of each bomb. We define $T_e = \min(T_1, T_2)$, a random variable for the time taken for either bomb to explode, and $T_b = \max(T_1, T_2)$, a random variable for the time taken for both bombs to explode. Find the pdf $p_e(t)$ corresponding to $T_e$, and $p_b(t)$ corresponding to $T_b$, satisfying

$$\int_0^t p_e(x)\,dx = P(T_e \leq t) \qquad \int_0^t p_b(x)\,dx = P(T_b \leq t)$$

Plot $p(t) = e^{-t}$ vs. $p_b(t)$ vs. $p_e(t)$. How long, on average, before any bomb explodes? How long, on average, before both bombs explode?

(Hint: Bayes rule is not useful here.)

**Solution.** We first look at the probability $P(T_b \leq t) = P(\max(T_1, T_2) \leq t)$.
Clearly, $P(\max(T_1, T_2) \leq t) = P(T_1 \leq t, T_2 \leq t)$ (as if $t$ is bigger than the maximum, it must be bigger than both.) Since the bomb's detonation times $T_1, T_2$ were sampled independently, and have the same distribution as $T$, then

$$P(T_1 \leq t, T_2 \leq t) = P(T_1 \leq t)P(T_2 \leq t).$$

Since both bombs have the same underlying distribution,

$$P(T_1 \leq t)P(T_2 \leq t) = P(T \leq t)^2 = \left(\int_0^t e^{-t}\,dt\right)^2 = (1 - e^{-t})^2$$

Then, by definition of the pdf, we have that

$$p_b(t) = \frac{d}{dt}P(\max(T_1, T_2) \leq t) = \frac{d}{dt}(1 - e^{-t})^2 = 2(1 - e^{-t})e^{-t}$$

for which the average explosion time is

$$\mathbb{E}_T[p_b(t)] = \int_0^\infty t p_e(t)\,dt = \int_0^\infty t \times 2(1 - e^{-t})e^{-t}\,dt = 1.5 \text{ minutes}$$

Now, we look at the probability $P(T_e \leq t) = P(\min(T_1, T_2) \leq t)$.
Clearly, $P(\min(T_1, T_2) \leq t) = P(T_1 \leq t \vee T_2 \leq t)$ (as if $t$ is less than the minimum, it has to be less than at least one.) We use the fact that $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$, and obtain

$$\begin{aligned}
&P(T_1 \leq t \vee T_2 \leq t) \\
&= P(T_1 \leq t) + P(T_2 \leq t) - P(T_1 \leq t, T_2 \leq t) \\
&= P(T_1 \leq t) + P(T_2 \leq t) - P(T_1 \leq t)P(T_2 \leq t) \\
&= 2P(T \leq t) - P(T \leq t)^2 \\
&= 2\int_0^t e^{-t}\,dt - \left(\int_0^t e^{-t}\,dt\right) \\
&= 2(1 - e^{-t}) - (1 - e^{-t})^2 \\
&= 2 - 2e^{-t} - (1 - 2e^{-t} + e^{-2t}) \\
&= 1 - e^{-2t}
\end{aligned}$$

Then by definition of pdf, we have that

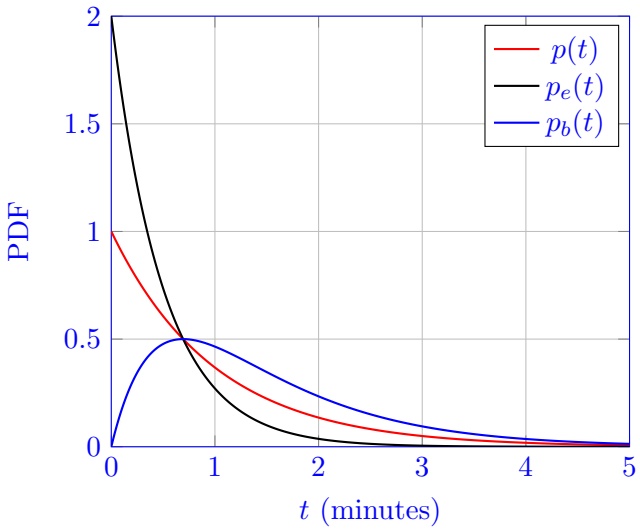
$$p_e(t) = \frac{d}{dt}P(\min(T_1, T_2) \leq t) = \frac{d}{dt}(1 - e^{-2t}) = 2e^{-2t}$$

[1] The bombs are placed sufficiently far apart that if one explodes, the other will be undisturbed. The detonation time of each bomb are independent of each other.

for which the average explosion time is

$$\mathbb{E}_X[p_e(t)] = \int_0^\infty t p_e(t)\, dt = \int_0^\infty 2te^{-2t}\, dt = 0.5 \text{ minutes}$$

Plotting the results,




**Question 3** **Definitions of variance**

Recall that given a continuous random variable $X$ defined over a domain $D \subset \mathbb{R}$ with probability distribution function $p(x) : D \to \mathbb{R}$, and a function $f(x) : \mathbb{R} \to \mathbb{R}$, we define [2]

$$\mathbb{E}_X[f(x)] := \int_D f(x)p(x)\, dx$$

The variance $\mathbb{V}[X]$ of a random variable $X$ is defined as

$$\mathbb{V}_X[x] := \mathbb{E}_X\left[(x - \mathbb{E}_X[x])^2\right]$$

It can also be represented in the alternate form

$$\mathbb{V}_X[x] := \mathbb{E}_X\left[x^2\right] - (\mathbb{E}_X[x])^2$$

Prove this!

**Solution.**

$$\mathbb{V}_X[x] = \int_D (x - \mathbb{E}_X[x])^2\, p(x)dx$$
$$= \int_D \left(x^2 - 2x\mathbb{E}_x[x] + \mathbb{E}_X[x]^2\right) p(x)dx$$
$$= \int_D x^2 p(x)dx - 2\mathbb{E}_X[x] \int_D xp(x)dx + \mathbb{E}_X[x]^2 \int_D p(x)dx$$

Note that $\int_D p(x)\, dx = 1$ by the definition of a probability distrubtion function.

$$= E_X\left[x^2\right] - 2\mathbb{E}_X[x]\mathbb{E}_X[x] + \mathbb{E}_X[x]^2$$
$$= \mathbb{E}_X\left[x^2\right] - \mathbb{E}_X[x]^2$$

[2] Note that $\int_D$ means to integrate over the entire domain $D$. For example, if $D = [0, 1]$, then $\int_D$ means the same thing as $\int_0^1$.

**Question 4** **Substitution of Random variables**

Assume that we have a random variable $X$ on the interval [0,1] characterized by a pdf $p(x) = \frac{3}{2}\sqrt{x}$. Let $Y$ be a random variable on [0,1] such that $Y = X^3$. Compute the pdf of $Y$

**Solution.** We compute the CDF function of $Y$,

$$F_Y(y) = P(Y \leq y) = P\left(X^3 \leq y\right) = P\left(X \leq y^{1/3}\right) = \int_0^{y^{1/3}} \frac{3}{2}\sqrt{x}dx = x^{3/2}\Big]_0^{y^{1/3}} = \sqrt{y}$$

We can obtain the pdf of $Y$ by differentiating.

$$p(y) = \frac{d}{dy}\sqrt{y} = \frac{1}{2\sqrt{y}}$$