The Australian National University
School of Computing

Semester 2, 2023
Theory Assignment 3 of 4

COMP3670/6670: Introduction to Machine Learning

Release Date. Sept 21, 2023
Due Date. 11:59 pm, Oct 9, 2023
Maximum credit. 100

Exercise 1 — Probability rules — (5 + 5 + 15 credits)

John and Ashley participated in an Easter Egg Hunt. After the hunt, John had a bag with 20 eggs, 15 red and 5 blue, and Ashley was able to collect 15 eggs, 10 red and 5 blue.

1. Without looking, John picks an egg from his bag. What is the probability that the egg is red?

   Solution. $\frac{3}{4}$.

2. Assuming we don't know the colour of the selected egg in 1. and don't return that egg to John's bag, he picks another egg from his bag. What is the probability that the second pick is a red egg?

   Solution. $p(2R) = p(2R|1R)p(1R) + p(2R|1B)P(1B) = \frac{14}{19} \cdot \frac{15}{20} + \frac{15}{19} \cdot \frac{5}{20} = \frac{3}{4}$

3. Now John put those two eggs back into his bag. When they got home, their dad Papi accidentally mixed the two bags into one. Papi picks two eggs from the big bag, sequentially. The first is a red egg and the second is blue, and gives them to Ashley. Assuming John and Ashley didn't eat any on their way home, what is the probability that those two eggs Papi picked are actually from Ashley's original bag?

   Solution. Denote the event the first Red egg belongs to Ashley as $1RA$, and the second Blue egg belongs to Ashley as $2BA$. These two events are independent, as picking a red won't affect the probability of the blue. Thus,

$$p(1RA, 2BA) = p(1RA)p(2BA) = \frac{10}{25} \cdot \frac{5}{10} = \frac{1}{5}$$

Exercise 2 — Geometric distributions and Bayes' rule — (5 + 5 + 10 + 10 credits)

Young statistician Terry is in a bar and sees the bartender start tossing a coin when each customer orders, and if the coin lands on a head, the customer will get a free drink.

1. Terry noted the first free drink was given out after 5 tosses. What is the probability of this, assuming the probability of the coin landing heads is p?

   Solution. This means first 4 tosses are tails, the 5th toss is a head. So $(1-p)^4 p$.

2. Terry's friend, Tao, is the tenth customer. Assuming no free drink was given between the fifth customer and Tao, what is the probability that Tao will get a free drink, given the first free drink? Comment on the result.

   Solution. This simply means the 6th to 9th are not free, and the 10th is free. So the probability of this is also $(1-p)^4 p$. This is because the Geometric distribution is "reset" after observing a success event (here, success means free drink).

3. Tao actually got a free drink! Terry estimates p by finding p_{ml} that maximises the probability of the observed free drink and non-free drink events. Find p_{ml}.

   Solution. The likelihood is given as $p^2(1-p)^8$. It's gradient over p is $2p(1-p)^8 - 8p^2(1-p)^7$. p by definition cannot be 0, so setting the gradient to 0, we have $p_{ml} = 0.2$.

4. Terry is not happy about this estimate. He then tries a Bayesian approach, by first placing a *beta* prior over p. The *beta* distribution has the following pdf,

$$f(p;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

   where α and β are two parameters of the distribution, and $B(\alpha,\beta)$ is the Beta function (you don't need to know the exact form to complete this exercise). For $\alpha > 1$ and $\beta > 1$, the mode of the distribution where the density is peaked is $\frac{\alpha-1}{\alpha+\beta-2}$. Assume $\alpha = \beta = 2$, find the posterior distribution, that is the distribution over p conditioned on the observed events, and then find the mode of this distribution and compare the posterior mode with the estimate in 3. Comment on the difference.

   Solution. What we have is the prior, $P(p) = \frac{1}{B(2,2)} \cdot p \cdot (1-p)$. The likelihood is given as $P(o|p) = p^2(1-p)^8$. The posterior update is given as

$$P(p|o) = \frac{P(p)P(o|p)}{P(o)} \propto P(p)P(o|p) = p^3(1-p)^9$$

   Either using the property of the conjugate prior or exact derivation leads to

$$P(p|o) = \frac{1}{B(4,10)} p^3(1-p)^9 = 2860 p^3(1-p)^9$$

   We calculate the MAP by letting the gradient of the posterior to 0:

$$3p^2(1-p)^9 - 9p^3(1-p)^8 = 0$$

   p cannot be 0, so $p_{map} = 0.25$. p_{map} is the mode we desire. The difference comes from the initial guess of α and β. In the bayesian setting, we initially assume the coin is fair by guessing $\alpha = \beta = 2$. In MLE setting, there is no such initial guess.

Exercise 3 **Gaussian distributions and Bayes' rule** (10 + 5 + 5 credits)

An explosion was detected by two sensors. Each sensor is only able to output a noisy estimate of the location of the explosion due to measurement noise. Assuming the two sensor outputs are y_1 and y_2, and

the likelihood of the exact location x given the sensor outputs is,

$$p(y_1|x)p(y_2|x) = \mathcal{N}(y_1; x, \sigma_1^2)\mathcal{N}(y_2; x, \sigma_2^2),$$

where σ_1^2 and σ_2^2 are the measurement noise variances. Assuming a prior distribution over the location $p(x) = \mathcal{N}(x; 0, \sigma_0^2)$, where σ_0^2 is the prior variance.

1. Find the posterior distribution $p(x|y_1, y_2)$

   Solution. Here the two sensors are independent of each other. By Bayes rule:

   $$p(x|y_1, y_2) = \frac{p(x)p(y_1, y_2|x)}{p(y_1, y_2)} = \frac{p(x)p(y_1|x)p(y_2|x)}{p(y_1, y_2)}$$

   $p(x|y_1, y_2)$ is guaranteed to be a Gaussian distribution. So here we only care about the characterising function of a Gaussian, which is $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Hence,

   $$p(x|y_1, y_2) \propto e^{-\frac{x^2}{2\sigma_0^2}} e^{-\frac{(y_1-x)^2}{2\sigma_1^2}} e^{-\frac{(y_2-x)^2}{2\sigma_2^2}} = e^{-\frac{x^2}{2\sigma_0^2} - \frac{(y_1-x)^2}{2\sigma_1^2} - \frac{(y_2-x)^2}{2\sigma_2^2}}$$

   Then we write it into something like

   $$e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

   By simplification,

   $$e^{-\frac{x^2}{2\sigma_0^2} - \frac{(y_1-x)^2}{2\sigma_1^2} - \frac{(y_2-x)^2}{2\sigma_2^2}} = C \cdot \exp\left\{ -\frac{(\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2)\left(x - \frac{y_1\sigma_0^2\sigma_2^2 + y_2\sigma_0^2\sigma_1^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2}\right)^2}{2\sigma_0^2\sigma_1^2\sigma_2^2} \right\}$$

   where C is a constant. This means

   $$p(x|y_1, y_2) \propto \exp\left\{ -\frac{(\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2)\left(x - \frac{y_1\sigma_0^2\sigma_2^2 + y_2\sigma_0^2\sigma_1^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2}\right)^2}{2\sigma_0^2\sigma_1^2\sigma_2^2} \right\}$$

   By simple comparison, we have

   $$\mu = \frac{y_1\sigma_0^2\sigma_2^2 + y_2\sigma_0^2\sigma_1^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2} \qquad \sigma^2 = \frac{\sigma_0^2\sigma_1^2\sigma_2^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2}$$

   And the posterior distribution is given by $\mathcal{N}(x; \mu, \sigma)$.

2. What happens to the posterior distribution

   (a) when the measurement noise variances are very large, $\sigma_1^2, \sigma_2^2 \to \infty$,

   Solution. For the variance,

   $$\frac{\sigma_0^2\sigma_1^2\sigma_2^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2} = \frac{\sigma_0^2}{\frac{\sigma_0^2}{\sigma_2^2} + \frac{\sigma_0^2}{\sigma_1^2} + 1} \to \sigma_0^2$$

   For the mean,

   $$\frac{y_1\sigma_0^2\sigma_2^2 + y_2\sigma_0^2\sigma_1^2}{\sigma_0^2\sigma_1^2 + \sigma_0^2\sigma_2^2 + \sigma_1^2\sigma_2^2} = \frac{y_1\frac{1}{\sigma_1^2} + y_2\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}} \to 0$$

(b) when the prior variance $\sigma_0^2 \to \infty$, and $\sigma_1 = \sigma_2$.

Solution. If $\sigma_1 = \sigma_2$,

$$\mu = \frac{(y_1 + y_2)\frac{1}{\sigma_1^2}}{\frac{2}{\sigma_1^2} + \frac{1}{\sigma_0^2}} \to \frac{y_1 + y_2}{2}$$

$$\sigma^2 = \frac{1}{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}} \to \frac{\sigma_1^2}{2}$$

Exercise 4 **Conjugate priors** (10 + 15 credits)

In the lecture, we discussed the Bayesian linear regression model when we assume

1. the likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is a Gaussian distribution $\mathcal{N}(\mathbf{y}; \boldsymbol{\Phi\theta}, \sigma^2\mathbf{I})$
2. the prior weight $\boldsymbol{\theta}$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}; \mathbf{m}_0, \mathbf{S}_0)$

Then we derive the posterior distribution. Note the order here: we first fix the likelihood, then we choose the prior distribution. So you may ask: will the choice of likelihood affect the choice of the prior? The answer is **yes**. In the example given, we fix the prior Gaussian, given a Gaussian likelihood **with known covariance**, the posterior is also a Gaussian. Given a likelihood function, the posterior will maintain the same probability distribution family as the **conjugate prior** after the Bayesian update. Here, the conjugate prior of a **Gaussian with known covariance** is another Gaussian distribution.

The advantage is obvious by assuming a Gaussian likelihood, as the conjugate prior and the posterior are all Gaussians, significantly simplifying the Bayesian analysis. However, in some cases, the likelihood can be a distribution other than the Gaussian, or a **Gaussian with unknown mean and covariance**. How should we set up the priors in these situations? We consider some one dimensional cases below:

1. Given that the Gamma distribution likelihood

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

where α is a positive constant and $\beta > 0$ is unknown. We place a Gamma prior over β, $\text{Gamma}(\beta; \alpha_0, \beta_0)$ as the conjugate prior. Prove the posterior distribution can be written as $\text{Gamma}(\beta; \alpha_0 + \alpha, \beta_0 + y)$.

Solution. The question is asking what is $p(\beta|y)$. Using Bayes rule,

$$\begin{aligned} p(\beta|y) \propto p(y|\beta)p(\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \beta^{\alpha_0-1} e^{-\beta_0\beta} \\ &\propto \beta^\alpha e^{-\beta y} \beta^{\alpha_0-1} e^{-\beta_0\beta} \\ &= \beta^{\alpha+\alpha_0-1} e^{-\beta(y+\beta_0)} \end{aligned}$$

As the posterior is guaranteed to also be a Gamma distribution, by fixing $\beta^{\alpha+\alpha_0-1}e^{-\beta(y+\beta_0)}$ we can find the distribution parameters. Hence, we have $\alpha' = \alpha + \alpha_0$, $\beta' = y + \beta_0$.

2. Given that the Gaussian distribution likelihood

$$p(y|\mu,\tau) = \mathcal{N}(y;\mu,\tau^{-1}) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2\tau}{2}}$$

where μ, τ are unknown and has the Normal-Gamma distribution

$$\text{NormalGamma}(\mu,\tau;\mu_0,\tau_0,\alpha_0,\beta_0) = \frac{\beta_0^{\alpha_0}\sqrt{\tau_0}}{\Gamma(\alpha_0)\sqrt{2\pi}} \tau^{\alpha_0-\frac{1}{2}} e^{-\beta_0\tau} e^{-\frac{\tau\tau_0(\mu-\mu_0)^2}{2}}$$

as the conjugate prior. Prove the posterior distribution can be written as

$$\text{NormalGamma}(\mu,\tau;\mu',\tau',\alpha',\beta')$$

And represent $\mu', \tau', \alpha', \beta'$ using $\mu_0, \tau_0, \alpha_0, \beta_0, y$.

Solution. The question is asking what is $p(\mu,\tau|y)$. Using Bayes rule,

$$\begin{aligned}
p(\mu,\tau|y) &\propto p(y|\mu,\tau)p(\mu,\tau) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2\tau}{2}} \frac{\beta_0^{\alpha_0}\sqrt{\tau_0}}{\Gamma(\alpha_0)\sqrt{2\pi}} \tau^{\alpha_0-\frac{1}{2}} e^{-\beta_0\tau} e^{-\frac{\tau\tau_0(\mu-\mu_0)^2}{2}} \\
&\propto e^{-\frac{(y-\mu)^2\tau}{2}} \tau^{\alpha_0} e^{-\beta_0\tau} e^{-\frac{\tau\tau_0(\mu-\mu_0)^2}{2}} \\
&= \tau^{\alpha_0} \exp\left\{-\frac{1}{2}\left[\tau(\tau_0+1)(\mu - \frac{y+\tau_0\mu_0}{\tau_0+1})^2 - \frac{(y+\tau_0\mu_0)^2}{\tau_0+1}\tau + y^2\tau + 2\beta_0\tau + \tau\tau_0\mu_0^2\right]\right\} \\
&= \tau^{\alpha_0} \exp\left\{-\frac{1}{2}\left[y^2 + 2\beta_0 + \tau_0\mu_0^2 - \frac{(y+\tau_0\mu_0)^2}{\tau_0+1}\right]\tau\right\} \exp\left\{-\frac{1}{2}\tau(\tau_0+1)(\mu - \frac{y+\tau_0\mu_0}{\tau_0+1})^2\right\} \\
&= \tau^{\alpha_0} \exp\left\{-\left[\frac{\tau_0}{\tau_0+1}\frac{(y-\mu_0)^2}{2} + \beta_0\right]\tau\right\} \exp\left\{-\frac{1}{2}\tau(\tau_0+1)(\mu - \frac{y+\tau_0\mu_0}{\tau_0+1})^2\right\}
\end{aligned}$$

We know the posterior is also a Normal-Gamma distribution. Hence by having the above formula, we can read off the updated parameters

$$\begin{aligned}
\alpha' &= \alpha_0 + \frac{1}{2} \\
\tau' &= \tau_0 + 1 \\
\mu' &= \frac{y+\tau_0\mu_0}{\tau_0+1} \\
\beta' &= \frac{\tau_0}{\tau_0+1}\frac{(y-\mu_0)^2}{2} + \beta_0
\end{aligned}$$