Assignment 1: Gaussians, Categories, and Clusters
Human and Machine Learning
Chiba Institute of Technology, School of Design & Science
Prof. Joseph Austerweil
Due Fri Jun 5, 2026 at 8:00pm

Many of the problems require making graphs. Please do not forget to label your axes and title the graph! You may plot multiple distributions on the same graph **as long as each distribution is clearly legible**.

1. Gaussians: In class, we derived (or will derive) the posterior and predictive distributions for a data point generated from a Gaussian-Gaussian model: having a Gaussian likelihood with unknown mean and known variance, and with a Gaussian prior on the mean of the likelihood with known mean and known variance. This model can be written in generative process notation[1] as:

$$\mu \sim \mathrm{N}(\mu_0, \sigma_0^2) \qquad\qquad x_1, \ldots, x_N | \mu, \sigma_x^2 \overset{iid}{\sim} \mathrm{N}(\mu, \sigma_x^2)$$

Remember that *iid* means *independent* and *identically distributed* and the generative process notation $x|\mu, \sigma_x^2 \sim \mathrm{N}(\mu, \sigma_x^2)$ means that given the values of parameters μ and σ_x^2, x is normally distributed with mean μ and variance σ_x^2. So, this means

$$p(x|\mu, \sigma_x^2) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{1}{2\sigma_x^2}(x-\mu)^2}$$

Note that it is traditional to use a zero subscript for the parameters for a prior distribution. For this model, there is a closed form solution for the posterior probability, $p(\mu|x_1, \ldots, x_N)$, and predictive probability, $p(x_{N+1}|x_1, \ldots, x_N)$. They are are:[2]

$$\text{Posterior :} \qquad \mu|x_1, \ldots, x_N \sim \mathrm{N}\left(\frac{\mu_0\sigma_0^{-2} + \sigma_x^{-2}\sum_{n=1}^N x_n}{\sigma_0^{-2} + N\sigma_x^{-2}}, \left[\sigma_0^{-2} + N\sigma_x^{-2}\right]^{-1}\right)$$

$$\text{Prediction} \quad x_{N+1}|x_1, \ldots, x_N \sim \mathrm{N}\left(\frac{\mu_0\sigma_0^{-2} + \sigma_x^{-2}\sum_{n=1}^N x_n}{\sigma_0^{-2} + N\sigma_x^{-2}}, \left[\sigma_0^{-2} + N\sigma_x^{-2}\right]^{-1} + \sigma_x^2\right)$$

So, the predictive distribution has the same mean as the posterior distribution, but it has larger variance (it is σ_x^2 larger). For this problem, use $\mu_0 = 0$ and $\sigma_0^2 = 1$. In this problem, we will explore how the number of data points and variance of the likelihood affect the posterior and predictive distributions.

(a) *Prior.* To provide a baseline, turn in a plot of the prior distribution. Please make sure your plot captures the "interesting" part of the distribution (i.e., the two extrema of the x-axis are the tails and the width and maximum of the bell are clearly visible).

[1] I try to use the following convention for variables: scalar variables are lowercase and italics (e.g., x), indices tend to be n, t, c or k, the largest index is uppercase (e.g., N), vector variables are bold and lowercase (e.g., $\mathbf{x}$), and matrix variables are bold and uppercase (e.g., $\mathbf{X}$).

[2] Remember that you can always look up these special models — whose posterior distribution is the same form as the prior distribution — on Wikipedia: https://en.wikipedia.org/wiki/Conjugate_prior.

(b) *One datum update:* Calculate and plot the posterior and predictive distributions after observing $x_1 = 2$ for $\sigma_x^2 = 0.25$ and $\sigma_x^2 = 4$ (that is 4 different distributions: the posterior and predictive for $\sigma_x^2 = 0.25$ and the posterior and predictive for $\sigma_x^2 = 4$). How does changing the variance of the likelihood affect the distributions? Are there any differences? Why?

(c) *Multiple data update:* Calculate and plot the posterior and predictive distributions given $(x_1, \ldots, x_5) = (2.1, 2.5, 1.4, 2.2, 1.8)$ for $\sigma_x^2 = 0.25$ and $\sigma_x^2 = 4$. How does this compare to the previous example? Note that the average of the data points is 2, and so both contribute the same average value. For cases that differ, why do they differ then? For those that do not, why don't they differ?

2. Categories In this problem, we will investigate how to make categorization decisions for two categories, where each is defined as a Gaussian distribution. First, you will derive the probability of an item being in one category or another. Then, we will explore how the variances and prior probability of each category affect the posterior and predictive distributions.

For this problem, we will assume that data are generated by first picking which of two categories $c = 1, 2$ it belongs to (according to their prior probability) and then generating the datum according to the corresponding category's likelihood. This results in the following generative process:

$$c_n|\theta \sim \text{Bernoulli}(\theta) \qquad x_n|\mu_{c(n)}, \sigma^2_{c(n)} \overset{iid}{\sim} \text{N}(\mu_{c(n)}, \sigma^2_{c(n)})$$

Note that $c(n)$ is the same as c_n. I use parentheses rather than a subscript to avoid having something that is "double subscripted.". In generative process notation, $c_n|\theta \sim \text{Bernoulli}(\theta)$ means that c_n, the category for data point n, is a Bernoulli random variable with parameter θ. This means c_n will be 1 with probability θ (and so with probability $1-\theta$, $c_n = 2$). So, the prior probability of category 1 is θ ($P(c_n = 1) = \theta$). For all of Problem 2, assume $\mu_1 = -1$ and $\mu_2 = 1$.

(a) *Derivation — Categorization.* Using Bayes' rule, derive the probability of a single datum being in category 1: $P(c_1 = 1|x_1)$.[3] You can assume that the values of μ_1, μ_2, σ_1^2, and σ_2^2 are given parameters (I didn't put them to the right of the | to save space). Show your work (if you do not know how to create equations on a computer, you can scan your handwritten derivation and include it as an image into your document). As the next problem depends on this answer, I will give you what your derivation should end up with. It is

$$P(c = 1|x_1) = \frac{\theta \text{N}(x_1; \mu_1, \sigma_1^2)}{\theta \text{N}(x_1; \mu_1, \sigma_1^2) + (1-\theta)\text{N}(x_1; \mu_2, \sigma_2^2)}$$

where $N(x; \mu, \sigma^2)$ is the probability density of x from a Normal distribution with mean μ and variance σ^2.

(b) *Categorization.* Calculate and plot the probability of being in category 1 (so, the x-axis is the x_1 value and the y-axis is $P(c_1 = 1|x_1)$) for $\theta = 0.5$ and for

[3] If you are wondering why P is uppercase here and lowercase in other cases, it has to do with whether it is a probability mass (uppercase) or density (lowercase) function. You do not have to worry about this distinction (I mess it up sometimes too!), but as a general rule, you should use the uppercase $P(\cdot)$ when it is a discrete random variable and the lowercase $p(\cdot)$ when it is a continuous random variable.

Variable	Meaning
x	A data point
c_n or $c(n)$	Category associated with data point n
$X \sim P(X = x)$	Random variable X has probability distribution $P(X = x)$ (or X is distributed P).
$\mathrm{N}(\mu, \sigma^2)$	The Normal distribution with mean μ and variance σ^2
$\mathrm{N}(x; \mu, \sigma^2)$	The probability of x under a Normal distribution with mean μ and variance σ^2.
Bernoulli(θ)	The Bernoulli distribution. It's a fancy way of saying the random variable is a biased coin flip, where the probability of heads is given by the parameter θ. When used in a generative process, such as $c \sim$ Bernoulli(θ), it means that c takes value 1 with probability θ and value 2 with probability $1 - \theta$.
θ	For the purposes of Problem 2, θ is the prior probability of category 1, $P(c_1)$.

Table 1: The different symbols and variables in Problem 2 and their meaning.

$\theta = 0.75$ with $\sigma_1^2 = \sigma_2^2 = 1$ (remember $\mu_1 = -1$ and $\mu_2 = 1$). Next, calculate and plot the probability of being in category 1 for $\theta = 0.5$ when $\sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$. Do this again, but for $\theta = 0.75$. Please make sure all of your plots show all of the interesting behavior (so make sure the range of your x- and y-axes are appropriate). Describe the effect of changing the prior and the variance on categorization decisions. What is the effect of varying them? Do they have the same effect? Why do they or why do they not?

(c) *Derivation — Prediction.* Using Bayes' rule and the *Law of Total Probability*, derive the probability of a data point $p(x)$ according to this model (note, this is without any given data). As the next problem depends on this answer, I will give you what your derivation should end up with. It is

$$p(x_1) = \theta \mathrm{N}(x_1; \mu_1, \sigma_1^2) + (1 - \theta)\mathrm{N}(x_1; \mu_2, \sigma_2^2)$$

(d) *Prediction.* As before, plot $p(x_1)$ for $\theta = 0.5$, and then for $\theta = 0.75$ with $\sigma_1^2 = \sigma_2^2 = 1$. Next, calculate and plot $P(x_1)$ for $\theta = 0.5$ when $\sigma_1^2 = 0.5$ and $\sigma_2^2 = 2$. Do this again, but for $\theta = 0.75$. How does the prior and variance of the likelihood affect $p(x_1)$? Please make sure all of your plots show all of the interesting behavior (so make sure the range of your x- and y-axes are appropriate). Describe the effect of changing the prior and the variance on $p(x_1)$. What is the effect of varying them? Do they have the same effect? Why do they or why do they not? Note that $p(x_1)$ is sometimes called the *marginal data distribution* and this type of model is called a *mixture model* because it composes a new probability distribution by"mixing" two (or more) distributions together.