MLPR Tutorial[1] Sheet 3

Reminders: Attempt the tutorial questions, and ideally discuss them before your tutorial — for instance at ML-Base. You can seek clarifications and hints on Hypothesis. Full answers will be released.

---

1. **A Gaussian classifier:**

   A training set consists of one-dimensional examples from two classes. The training examples from class 1 are $\{0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25\}$ and the examples from class 2 are $\{0.9, 0.8, 0.75, 1.0\}$.

   a) Fit a one-dimensional Gaussian to each class by matching the mean and variance. Also estimate the class probabilities π_1 and π_2 by matching the observed class fractions. (This procedure fits the model with maximum likelihood: it selects the parameters that give the training data the highest probability.) Sketch a plot of the scores $p(x, y) = P(y)\, p(x \mid y)$ for each class y, as functions of input location x.

   b) What is the probability that the test point $x = 0.6$ belongs to class 1? Mark the decision boundary/ies on your sketch, the location(s) where $P(\text{class } 1 \mid x) = P(\text{class } 2 \mid x) = 0.5$. You are not required to calculate the location(s) exactly.

   c) Are the decisions that the model makes reasonable for very negative x and very positive x? Are there any changes we could consider making to the model if we wanted to change the model's asymptotic behaviour?

2. **More practice with Gaussians:**

   N noisy independent observations are made of an unknown scalar quantity m:

   $$x^{(n)} \sim \mathcal{N}(m, \sigma^2).$$

   a) We don't give you the raw data, $\{x^{(n)}\}$, but tell you the mean of the observations:

   $$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}.$$

   What is the likelihood[2] of m given only this mean $\bar{x}$? That is, what is $p(\bar{x} \mid m)$?[3]

   b) A *sufficient statistic* is a summary of some data that contains all of the information about a parameter.

      i) Show that $\bar{x}$ is a sufficient statistic of the observations for m, assuming we know the noise variance σ^2. That is, show that $p(m \mid \bar{x}) = p(m \mid \{x^{(n)}\}_{n=1}^{N})$.

      ii) **Optional part:** If we don't know the noise variance σ^2 or the mean, is $\bar{x}$ still a sufficient statistic in the sense that $p(m \mid \bar{x}) = p(m \mid \{x^{(n)}\}_{n=1}^{N})$? Explain your reasoning.

      Note: In i) we were implicitly conditioning on σ^2: $p(m \mid \bar{x}) = p(m \mid \bar{x}, \sigma^2)$. In this part, σ^2 is unknown, so $p(m \mid \bar{x}) = \int p(m, \sigma^2 \mid \bar{x})\, \mathrm{d}\sigma^2$. Although no detailed mathematical manipulation (or solving of integrals) is required.

---

1. Parts of this tutorial sheet are based on previous versions by Amos Storkey, Charles Sutton, and Chris Williams
2. We're using the traditional statistics usage of the word "likelihood": it's a function of parameters given data, equal to the *probability* of the data given the parameters. You should avoid saying "likelihood of the data" (Cf p29 of MacKay's textbook), although you'll see that usage too.
3. The sum of Gaussian outcomes is Gaussian distributed; you only need to identify a mean and variance.

3. **Bayesian regression with multiple data chunks:**

   This question involves simulations on a computer. You will generate datasets and calculate posterior distributions. If you have trouble, at least try to work out roughly what should happen for each part.

   We will use the probabilistic model for regression models from the lectures:

   $$p(y \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(y;\, f(\mathbf{x}; \mathbf{w}),\, \sigma_y^2).$$

   In this question, we set $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}$, where $\mathbf{w} = [w_1 \;\; w_2]^\top$ and $\mathbf{x} = [x_1 \;\; x_2]^\top$, and $\sigma_y^2 = 1^2 = 1$. We assume the following prior distribution:

   $$p(\mathbf{w}) = \mathcal{N}(\mathbf{w};\, [-5 \;\; 0]^\top,\, \sigma_{\mathbf{w}}^2 \mathbb{I}),$$

   where $\sigma_{\mathbf{w}}^2 = 2^2 = 4$.

   Generating data: Generate some synthetic dataset as below. Only generate the data once and do not change it when working on the parts of this question.

   First, draw a $\tilde{\mathbf{w}}$ from the prior distribution (i.e. draw a sample from the Gaussian). In the whole dataset generation process, draw this $\tilde{\mathbf{w}}$ only once and keep it constant.

   We will generate two chunks of data $\mathcal{D}_1$ and $\mathcal{D}_2$. For chunk $\mathcal{D}_1 = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{15}$, generate 15 input-output pairs where each $\mathbf{x}^{(n)}$ is drawn from $\mathcal{N}(\mathbf{x}^{(n)};\, \mathbf{0}, 0.5^2\mathbb{I})$. For chunk $\mathcal{D}_2 = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=16}^{45}$, generate 30 input-output pairs where each $\mathbf{x}^{(n)}$ is drawn from $\mathcal{N}(\mathbf{x}^{(n)};\, [0.5 \;\; 0.5]^\top, 0.1^2\mathbb{I})$. For both chunks, draw the $y^{(n)}$ outputs using the $p(y \mid \mathbf{x}, \mathbf{w})$ observation model above.

   Visualising distributions on the weights: For each part that asks you to visualise a distribution, draw 400 weight vectors from the distribution and show the samples in a 2D scatter plot. In each plot, show $\tilde{\mathbf{w}}$ as a bold cross.

   a) Visualise the prior distribution and discuss the relationship between the parameters and the belief expressed by the prior.

   b) Using equations from the lecture notes, calculate and visualise the posterior distributions that you get after observing:

      i. dataset $\mathcal{D}_1$, i.e. $p(\mathbf{w} \mid \mathcal{D}_1)$

      ii. dataset $\mathcal{D}_2$ but not $\mathcal{D}_1$, i.e. $p(\mathbf{w} \mid \mathcal{D}_2)$

      iii. datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, i.e. $p(\mathbf{w} \mid \mathcal{D}_1, \mathcal{D}_2)$

      iv. no observations, i.e. $p(\mathbf{w} \mid \{\})$

   c) Take the posterior that you got in b) for $p(\mathbf{w} \mid \mathcal{D}_1)$ and now use it as a prior. Calculate the new posterior that you get after observing $\mathcal{D}_2$. Numerically compare the mean and covariance with those of $p(\mathbf{w} \mid \mathcal{D}_1, \mathcal{D}_2)$.

   d) Is it important that the inputs were drawn from Gaussian distributions? What would happen if the $\mathbf{x}^{(n)}$ in $\mathcal{D}_2$ were drawn from a uniform distribution on the unit square?

   e) What distribution is $p(\mathbf{w} \mid \mathcal{D}_1)$ approaching when we let $\sigma_{\mathbf{w}}^2$ approach infinity?

   f) Now assume that our observations y are corrupted by additional Gaussian noise:

      $$p(z \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(z;\, y, \sigma_z^2).$$

      So we now observe datasets $\mathcal{D}_z = \{\mathbf{x}^{(n)}, z^{(n)}\}_{n=1}^{N}$. What is the new posterior distribution $p(\mathbf{w} \mid \mathcal{D}_z)$? *Hint:* Recall how you did Question 2.a) of this sheet.