Engineering Tripos Part IIA THIRD YEAR

Module 3F8: Inference

Solutions to Example Sheet 1: Introductory Inference Problems and Regression

Introductory Inference Problems

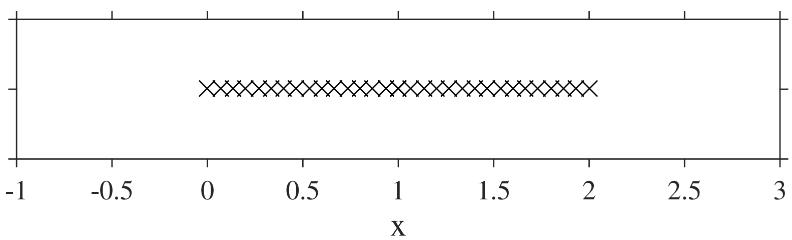
1. Maximum likelihood fitting of a Gaussian

   (a) Explain the terms likelihood function, prior probability distribution, and posterior probability distribution, in the context of the inference of parameters θ from data $\mathcal{D}$.

   (b) A random variable x is believed to have a probability distribution which is Gaussian with mean μ and standard deviation equal to 1,

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right).$$

   A sample of $N = 32$ data points is collected $\{x_n\}_{n=1}^N$ that are believed to be drawn independently from this distribution. The dataset is shown below:



   The first and second moments of these data are $\frac{1}{N}\sum_{n=1}^N x_n = 1$ and $\frac{1}{N}\sum_{n=1}^N x_n^2 = 1.3$.
   Sketch the likelihood as a function of μ for the dataset. Label the position of the maximum and its width. You do not need to compute the value of the likelihood at its maximum.

1. $p(x_n|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_n-\mu)^2}$

$$p(\{x_n\}_{n=1}^N|\mu) = \prod_n p(x_n|\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-\frac{1}{2}\sum_n (x_n-\mu)^2}$$

$$= (2\pi)^{-N/2} e^{-\frac{1}{2}\left(\sum_n x_n^2 - 2\mu\sum_n x_n + \mu^2 N\right)}$$

⇒ Gaussian form in μ

$$= (2\pi)^{-N/2} e^{-\frac{1}{2}N\left(\langle x^2\rangle_{p_0} - 2\mu\langle x\rangle_{p_0} + \mu^2\right)}$$

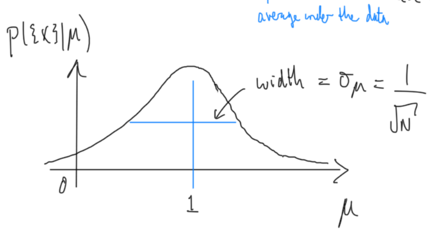
$$= Z\,\mathcal{N}(\mu;\mu_\mu,\sigma_\mu^2) = \frac{Z}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_\mu)^2}$$

$$= \frac{Z}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{1}{2\sigma_\mu^2}\left(\mu^2 - 2\mu\mu_\mu - \mu_\mu^2\right)}$$

compare
[complete the square]

$\Rightarrow \sigma_\mu^2 = \frac{1}{N} \qquad \mu_\mu = \langle x\rangle_{p_0} = 1$ (could also find Z, but question does not ask for it)

↑ average under the data



NB. when fitting a Gaussian, the likelihood only depend on the data's 1st & 2nd moments.
So even though the data appear to come from a uniform density here, we only need the two moments.

2. Inference in a Gaussian model

A noisy depth sensor measures the distance to an object an unknown distance d metres away. The depth can be assumed, *a priori*, to be distributed according to a standard Gaussian distribution $p(d) = \mathcal{N}(d; 0, 1)$. The depth sensor returns y a noisy measurement of the depth, that is also assumed to be Gaussian $p(y|d, \sigma_y^2) = \mathcal{N}(y; d, \sigma_y^2)$.

(a) Compute the posterior distribution over the depth given the observation, $p(d|y, \sigma_y^2)$.

(b) What happens to the posterior distribution as the measurement noise becomes very large $\sigma_y^2 \to \infty$? Comment on this result.

The formula for the probability density of a Gaussian distribution of mean μ and variance σ^2 is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

2. $p(d) = N(d; 0, 1)$ $\qquad$ $p(y|d, \sigma_y^2) = N(y; d, \sigma_y^2)$

a) $$p(d|y) \propto p(y|d,\sigma_y^2)\,p(d) \propto e^{-\frac{1}{2\sigma_y^2}(y-d)^2 - \frac{1}{2}d^2}$$

$$= e^{-\frac{1}{2}\left[d^2\left(\frac{1}{\sigma_y^2}+1\right) - \frac{2dy}{\sigma_y^2} + y^2/\sigma_y^2\right]}$$ $\Leftarrow$ Gaussian form

compare coefficients $\Downarrow$

$$= e^{-\frac{1}{2\sigma_{d|y}^2}(d-\mu_{d|y})^2} = e^{-\frac{1}{2}\left(\frac{d^2}{\sigma_{d|y}^2} - \frac{2d\mu_{d|y}}{\sigma_{d|y}^2} + \frac{\mu_{d|y}^2}{\sigma_{d|y}^2}\right)}$$

$$\therefore \quad \sigma_{d|y}^2 = \frac{1}{1/\sigma_y^2 + 1} = \frac{\sigma_y^2}{1+\sigma_y^2}$$

$$\mu_{d|y} = \sigma_{d|y}^2 \frac{y}{\sigma_y^2} = \frac{y}{1+\sigma_y^2}$$

b) $\sigma_y^2 \to \infty \quad \Rightarrow \quad \sigma_{d|y}^2 \to 1$ (this is the prior variance as it should be as now the sensor is so noisy it does not tell us anything over & above our a priori beliefs)

$\Rightarrow \mu_{d|y} \to 0$ (again collapses back to the prior)

Note also that when $\sigma_y^2 \to 0$ the sensor gives us perfect information about y so $\sigma_{d|y}^2 \to 0$ & $\mu_{d|y} \to y$ as expected

3. Bayesian inference for a biased coin*

A sequence of coin tosses are observed from a biased coin $x_{1:N} = \{0, 1, 1, 0, 1, 1, 1, 1, 0\}$ where $x_n = 1$ indicates flip n was a head and $x_n = 0$ indicates that it was tails. An experimenter would like to estimate the coin's probability of landing heads, ρ, from these data.

The experimenter assumes that the coin flips are drawn independently from a Bernoulli distribution $p(x_n|\rho) = \rho^{x_n}(1-\rho)^{1-x_n}$ and uses a prior distribution of the form

$$p(\rho|n_0, N_0) = \frac{1}{Z(n_0, N_0)}\rho^{n_0}(1-\rho)^{N_0-n_0}.$$

Here n_0 and N_0 are parameters set by the experimenter to encapsulate their prior beliefs. $Z(n_0, N_0)$ returns the normalising constant of the distribution as a function of the parameters, n_0 and N_0.

(a) Compute the posterior distribution over the bias $p(\rho|x_{1:N}, n_0, N_0)$.

(b) Compute the *maximum a posteriori* (MAP) estimate for the bias.

(c) Provide an intuitive interpretation for the parameters of the prior distribution, n_0 and N_0. For what setting of n_0 and N_0 does the MAP estimate become equal to the maximum-likelihood estimate?

3. a) $p(\rho \mid x_{1:N}, n_0, N_0) \propto p(\rho \mid n_0, N_0) \prod_{n=1}^{N} p(x_n \mid \rho)$

$$= \frac{1}{Z(n_0, N_0)} \rho^{n_0} (1-\rho)^{N_0 - n_0} \rho^{\sum_n x_n} (1-\rho)^{N - \sum_n x_n}$$

$$= \frac{1}{Z} \rho^{n_0 + n} (1-\rho)^{N_0 + N - n_0 - n}$$

where $n = \sum_n x_n$
i.e. # of 1s in dataset

$$\therefore \; p(\rho \mid x_{1:N}, n_0, N_0) = \frac{1}{Z(n', N')} \rho^{n'} (1-\rho)^{N' - n'}$$

where $n' = n_0 + n$
$N' = N_0 + N$

(This is a Beta distribution & it is conjugate to the likelihood, meaning the posterior has the same form)

b) $\log p(\rho \mid x_{1:N}, n_0, N_0) = -\log Z + n' \log \rho + (N' - n') \log(1-\rho)$

$$\frac{d}{d\rho} \log p(\rho_{MAP} \mid x_{1:N}, n_0, N_0) = \frac{n'}{\rho_{MAP}} - \frac{(N' - n')}{1 - \rho_{MAP}} = 0$$

$$(1 - \rho_{MAP}) n' - \rho_{MAP}(N' - n') = 0 \quad \Rightarrow \quad \rho_{MAP} = \frac{n'}{N'}$$

c) N_0 = # of pseudo data points seen before real data

n_0 = # of 1's in pseudo data

ML estimate is recovered when $N_0 = n_0 = 0 \Rightarrow$ flat prior distribution (no pseudo data)

4. Inferential game show*

   On a game show, a contestant is told the rules as follows:

   There are four doors, labelled 1, 2, 3 and 4. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other three doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2, 3 and 4, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.

   At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to one of the other closed doors. All the doors will then be opened and you will receive whatever is behind your final choice of door.

   (a) Imagine that the contestant chooses door 1 first; then the gameshow host opens door 4, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2 or 3, or (c) does it make no difference?

   (b) Use Bayes' rule to solve the problem.

4) WLG let door 1 be the one selected by the contestant

let S = position of prize $\quad S \in \{1,2,3,4\}$

A priori we assume $p(S=k) = 1/4$

The datum we receive after choosing door 1 is either $D=2, D=3, D=4$

ie doors 2, 3 or 4 are opened.

Assume that when the host has a choice about which door to open he selects uniformly between those doors not associated with the prize.

ie

$$
\begin{array}{llll}
p(D=2|S=1) = 1/3 & p(D=2|S=2) = 0 & p(D=2|S=3) = 1/2 & p(D=2|S=4) = 1/2 \\
p(D=3|S=1) = 1/3 & p(D=3|S=2) = 1/2 & p(D=3|S=3) = 0 & p(D=3|S=4) = 1/2 \\
p(D=4|S=1) = 1/3 & p(D=4|S=2) = 1/2 & p(D=4|S=3) = 1/2 & p(D=4|S=4) = 0
\end{array}
$$

Now apply Bayes' theorem:

$$p(S=k \mid D=4) = \frac{p(D=4 \mid S=k)\, p(S=k)}{p(D=4)}$$

$$p(S=1 \mid D=4) = \frac{\frac{1}{3} \cdot \frac{1}{4}}{p(D=4)} \qquad p(S=2 \mid D=4) = \frac{\frac{1}{2} \cdot \frac{1}{4}}{p(D=4)} \qquad p(S=3 \mid D=4) = \frac{\frac{1}{2} \cdot \frac{1}{4}}{p(D=4)} \qquad p(S=4 \mid D=4) = 0$$

$= 1/4$ ← same as prior $\qquad = 3/8$ ← greater than prior $\qquad = 3/8$ ← greater than prior

So, if we switch to doors 2 or 3 we will increase our chances of winning from $1/4$ to $3/8$ ie. 1.5x.

To get an intuition for the fact that the opening of the door by the host provides information, consider 100 doors & the host opening 98 of them.

This is a version of the Monty Hall problem (see pg 57 of David MacKay's information theory & inference book)

5. Bayesian decision theory*

A data-scientist has computed a complex posterior distribution over a variable of interest, x, given observed data y, that is $p(x|y)$. They would like to return a point estimate of x to their client. The client provides the data-scientist with a reward function $R(\hat{x}, x)$ that indicates their satisfaction with a point estimate $\hat{x}$ when the true state of the variable is x.

(a) Explain how to use *Bayesian Decision Theory* to determine the optimal point estimate, $\hat{x}$.

(b) Compute the optimal point estimate $\hat{x}$ in the case when the reward function is the negative square error between the point estimate and the true value, $R(\hat{x}, x) = -(\hat{x} - x)^2$. Comment on your result.

(c) Compute the optimal point estimate $\hat{x}$ in the case when the reward function is the negative absolute error between the point estimate and the true value, $R(\hat{x}, x) = -|\hat{x} - x|$. Comment on your result.

5)

a) $$\hat{x}_* = \arg\max_{\hat{x}} \int R(\hat{x}, x)\, p(x|y)\, dx$$

b) Find optimum: $$-\frac{d}{d\hat{x}} \int (x-\hat{x})^2 p(x|y)\, dx = 0$$

$$+2 \int (x - \hat{x}_*)\, p(x|y)\, dx = 0$$

$$\therefore \int x\, p(x|y)\, dx = \hat{x}_*$$

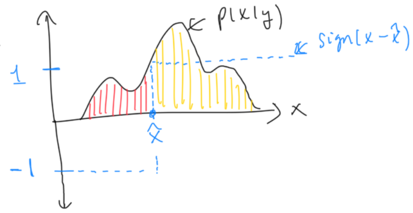
ie the posterior mean minimises the expected squared error

c) Find optimum: $$-\frac{d}{d\hat{x}} \int |x-\hat{x}|\, p(x|y)\, dx = 0$$

$$= -\frac{d}{d\hat{x}} \int \sqrt{(x-\hat{x})^2}\, p(x|y)\, dx$$

$$= +\frac{1}{\cancel{2}} \cdot \cancel{2}^{\,1} \cdot \int \underbrace{\frac{(x-\hat{x})}{\sqrt{(x-\hat{x})^2}}}_{\mathrm{sign}(x-\hat{x})} p(x|y)\, dx$$

Schematic of integral



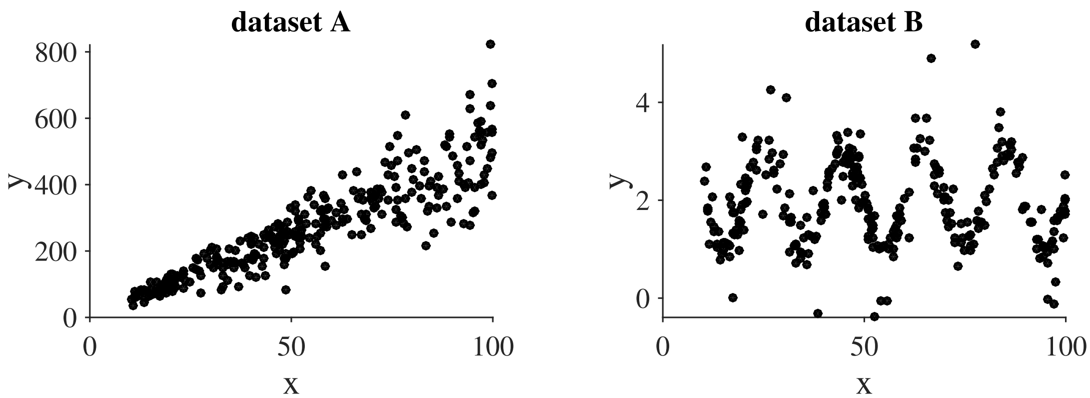
⇒ need to find the point where the red & yellow areas are equal

⇒ median of the distribution (has half the density above it & half below)

Regression

6. Probabilistic models for regression*

A machine learner observes two separate regression datasets comprising scalar inputs and outputs $\{x_n, y_n\}_{n=1}^N$ shown below.



(a) Suggest a suitable regression model, $p(y_n|x_n)$ for the dataset A. Indicate sensible settings for the parameters in your proposed model where possible. Explain your modelling choices.

(b) Suggest a suitable regression model, $p(y_n|x_n)$ for the dataset B. Indicate sensible settings for the parameters in your proposed model where possible. Explain your modelling choices.

a) Linear trend, rough gradient ≈ 5, intercept @ $\{0,0\}$

Noise appears Gaussian BUT standard deviation grows with x

$\therefore$ suggest $y_n(x) = 5x + \sigma(x)\,\varepsilon_n$ $\quad \varepsilon_n \sim N(0,1)$

where $\sigma(x) = |x|$

Many reasonable choices here, might be good to discuss what people have come up with in the supervision.

b) Sinusoidal trend, time period of rough 25 time steps heavy tailed noise (outliers)

amplitude is ≈ 1

$y_n(x) = 2 + \sin\left(\frac{2\pi}{25}x\right) + \varepsilon_n$ $\quad \varepsilon_n \sim$ Student-t

note mean is 2

heavy tailed, mean 0 variance ≈ 1 (hard to guess)

degree of freedom parameter ≈ 2.1 (even harder to guess)

Again it's a good one to discuss.

7. Maximum-likelihood learning for a simple regression model

Consider a regression problem where the data comprise N scalar inputs and outputs, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, and the goal is to predict y from x.

Assume a very simple linear model, $y_n = ax_n + \epsilon_n$, where the noise ϵ_n is Gaussian with zero mean and variance 1.

(a) Provide an expression for the log-likelihood of the parameter a.

(b) Compute the maximum likelihood estimate for a.

a) $$p(\{y_n\}_{n=1}^N \mid a, \{x_n\}_{n=1}^N) = \prod_{n=1}^N p(y_n \mid a, x_n)$$

$$\therefore \log p(\{y_n\}_{n=1}^N \mid a, \{x_n\}_{n=1}^N) = \sum_n \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}(y_n - ax_n)^2 \right]$$

$$= -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_n (y_n - ax_n)^2 = \mathcal{L}(a)$$

b) $$\left.\frac{d\mathcal{L}(a)}{da}\right|_{a_{ML}} = \sum_n x_n(y_n - a_{ML}x_n) = 0$$

$$\Rightarrow a_{ML} = \sum_n x_n y_n \Big/ \sum_n x_n^2$$

8. Maximum-likelihood learning for multi-output regression*

A data-scientist has collected a regression dataset comprising N scalar inputs ($\{x_n\}_{n=1}^N$) and N scalar outputs ($\{y_n\}_{n=1}^N$). Their goal is to predict y from x and they have assumed a very simple linear model, $y_n = ax_n + \epsilon_n$.

The data-scientist also has access to a second set of outputs ($\{z_n\}_{n=1}^N$) that are well described by the model $z_n = x_n + \epsilon'_n$.

The noise variables ϵ_n and ϵ'_n are known to be zero mean correlated Gaussian variables

$$p\left(\begin{bmatrix} \epsilon_n \\ \epsilon'_n \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \epsilon_n \\ \epsilon'_n \end{bmatrix}; \mathbf{0}, \Sigma\right) \quad \text{where} \quad \Sigma^{-1} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

(a) Provide an expression for the log-likelihood of the parameter a.

(b) Compute the maximum likelihood estimate for a.

(c) Do the additional outputs $\{z_n\}_{n=1}^N$ provide useful additional information for estimating a? Explain your reasoning.

The formula for the probability density of a multivariate Gaussian distribution of mean μ and covariance Σ is given by

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right).$$

a) $\mathcal{L}(a) = \log p(\{y_n\}_{n=1}^N, \{z_n\}_{n=1}^N \mid \{x_n\}_{n=1}^N, a) = \sum_n \log p(y_n, z_n \mid x_n, a)$

where $p(y_n, z_n \mid a, x_n) = \mathcal{N}\left(\begin{bmatrix} y_n \\ z_n \end{bmatrix}; \begin{bmatrix} ax_n \\ x_n \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}^{-1}\right)$

$$\therefore \mathcal{L}(a) = \sum_n -\frac{1}{2}\left(\begin{bmatrix} y_n \\ z_n \end{bmatrix} - \begin{bmatrix} ax_n \\ x_n \end{bmatrix}\right)^\top \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \left(\begin{bmatrix} y_n \\ z_n \end{bmatrix} - \begin{bmatrix} ax_n \\ x_n \end{bmatrix}\right) - \frac{N}{2}\log\det(2\pi\Sigma)$$

$$= -\frac{1}{2}\sum_n (y_n - ax_n)^2 - \frac{1}{2}\sum_n (y_n - ax_n)(z_n - x_n) - \frac{1}{2}\sum_n (z_n - x_n)^2 - \frac{N}{2}\log\det 2\pi\Sigma$$

bit from just observing ys

extra bit from observing zs

b) $$\frac{d\mathcal{L}(a)}{da} = \sum_n (y_n - ax_n)x_n + \frac{1}{2}\sum_n (z_n - x_n)x_n = 0$$

$$= \sum_n y_n x_n + \frac{1}{2}\sum_n z_n x_n - \frac{1}{2}\sum_n x_n^2 - a\sum_n x_n^2$$

$$\therefore \quad a = \Big(\sum_n y_n x_n + \underbrace{\frac{1}{2}\sum_n \overbrace{(z_n - x_n)}^{\epsilon'_n} x_n}_{\text{new contribution from observing zs}}\Big) \Big/ \sum_n x_n^2 \quad \text{(max likelihood estimate)}$$

c) The additional outputs change the ML estimate of a. This means that they must provide useful information about a. They do this because the noise in z_n is correlated with the noise in y_n & so observing z_n reveals information about the noise ϵ_n & allows more accurate identification of a.

Richard E. Turner