Tutorial 2:

Exercises for Block I

Exercise 2: Maximum Density

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to

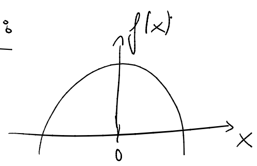
$$p_y(y) = p_x(g(y))|g'(y)| \tag{1}$$

i) By differentiating 1, show that the location $\hat{y}$ of the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable.

ii) Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

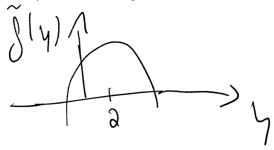
i) Normal functions:

- Take $f(x)$:



- Change to new variable y using $x = g(y)$. e.g.: $g(a) = a - 2$

$\hookrightarrow \tilde{f}(y) = f(g(y))$ $\qquad \hookrightarrow$



- Say $f(x)$ has a mode (maximum) at $\hat{x}$, s.t. $f'(\hat{x}) = 0$

$\hookrightarrow$ What is the mode $\hat{y}$ of $\tilde{f}(y)$?

- $\tilde{f}'(\hat{y}) \overset{!}{=} 0$

assuming $g'(\hat{y}) \neq 0$

$\Leftrightarrow f'(g(\hat{y})) \cdot g'(\hat{y}) = 0 \Leftrightarrow f'(g(\hat{y})) = 0$

$\hookrightarrow$ But we also know $f'(\hat{x}) = 0$

$\hookrightarrow \hat{x} = g(\hat{y})$

Density:

- Consider $p_X(x)$ with mode $\hat{x}$, change of variables: $x = g(y)$

$\hookrightarrow p_Y(y) = p_X(g(y)) \cdot |g'(y)|$

- Perform derivative w.r.t. $y \Rightarrow$ it must be 0 for $\hat{y}$!

$$\frac{d p_y(y)}{d y} = \frac{d p_x(g(y))}{d y} \cdot |g'(y)| + p_x(g(y)) \cdot \frac{d|g'(y)|}{dy}$$

$$= \frac{d p_x(g(y))}{d\,g(y)} \cdot \frac{d\,g(y)}{\ldots} \cdot |g'(y)| \ldots$$

$$= \frac{d p_x(g(y))}{d g(y)} \cdot \frac{d\, g(y)}{d y} \cdot |g'(y)| + \text{"}$$

↳ Since $\frac{dp_x(x)}{dx}\Big|_{x=\hat{x}} = 0$, if we chose $\hat{x} = g(\hat{y})$, then $\frac{dp_x(g(y))}{dg(y)}\Big|_{y=\hat{y}} = 0$.

↳ What about second term? → Doesn't go away!

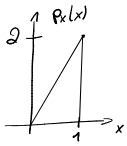
↳ $\underbrace{p_x(g(y)) \cdot \frac{d|g'(y)|}{dy}}$

ii) Look at this: If $g(y) = a \cdot y + b$, then $|g'(y)| = |a|$ and thus

$$\frac{d|g'(y)|}{dy} = 0 \implies \hat{x} = g(\hat{y})!$$

Example:

Let $p_x(x) = 2x$, $x \in [0,1]$:



↳ $\hat{x} = 1$

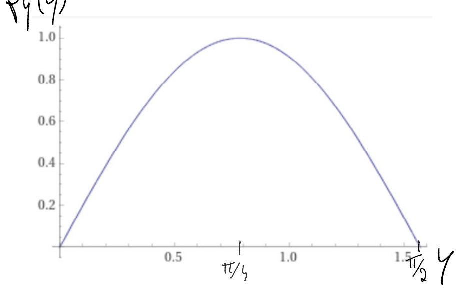
Now, set $x = \sin(y)$.

↳ $p_y(y) = 2 \cdot \sin(y) \cdot |\cos(y)|$, $y \in [0, \frac{\pi}{2}]$

$= 2 \sin(y) \cdot \cos(y)$

$= \sin(2y)$ $\quad \sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \sin(\beta)\cdot\cos(\alpha)$

Plot of $p_y(y)$:



↳ Obviously, $\hat{x} \neq \sin(\hat{y})$.

$D = \{x_i\}_{i=1}^N$

Exercise 7: Maximum likelihood estimates

Verify by setting the derivatives of the log likelihood

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$\begin{cases} p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ p(\{x_1,\dots,x_n\}|\mu,\sigma^2) = \prod_{i=1}^{n} p(x_n|\mu,\sigma^2) \end{cases}$$

with respect to μ and σ^2 equal to zero:

i) $\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$ (see [1] 1.55)

ii) $\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$ (see [1] 1.56)

i) $\partial\, \partial_\mu (\ln(X|\mu,\sigma^2))$ $\quad 1$

ii) $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$ (see [1] 1.56)

i) $$\frac{\partial \ln(p(X|\mu,\sigma^2))}{\partial \mu} = -\frac{1}{\sigma^2}\sum(x_n-\mu)\cdot(-1)$$

$$= \frac{1}{\sigma^2}\sum(x_n-\mu) \overset{!}{=} 0$$

$$\Leftrightarrow \quad \frac{1}{\sigma^2}\sum x_n = \frac{1}{\sigma^2}\cdot N\cdot\mu$$

$\cdot\,\sigma^2/N$

$$\Leftrightarrow \quad \mu_{ML} = \frac{1}{N}\sum x_n$$

$$\frac{-1}{2\sigma^2} \,\hat{=}\, -\frac{1}{2}(\sigma^2)^{-1}$$

ii) $$\frac{\partial \ln(p(X|\mu,\sigma^2))}{\partial \sigma^2} = \frac{1}{2}(\sigma^2)^{-2}\cdot\sum(x_n-\mu)^2 - \frac{N}{2}\cdot\frac{1}{\sigma^2} \overset{!}{=} 0$$

$\cdot\, 2(\sigma^2)^2/N$

$$\Leftrightarrow \quad \frac{1}{N}\sum(x_n-\mu)^2 - \sigma^2 \overset{!}{=} 0$$

set $\mu = \mu_{ML}$

$$\Leftrightarrow \quad \sigma^2_{ML} = \frac{1}{N}\sum(x_n-\mu_{ML})^2$$

Exercise 8: True variance

Suppose that the variance of a Gaussian is estimated using $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$ (see [1] 1.56) but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2 "unbiasedness"

$$\mathbb{E}_{\{x_n\}}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{x_n}\left[x_n^2 - 2x_n\mu + \mu^2\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\underbrace{(\mu^2+\sigma^2)}\, - 2\mu\mu + \mu^2\right)$$

$$= \sigma^2$$

For any two r.v.'s X, Y: $\mathbb{E}[XY] = \mathbb{E}[X]\cdot\mathbb{E}[Y] + \mathrm{Cov}(X,Y)$

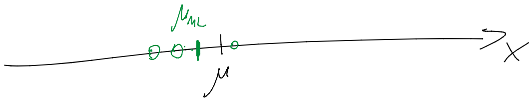
$\hookrightarrow \mathbb{E}[X^2] = \mu^2 + \sigma^2$

Bonus: What if we use μ_{ML}?

$$\hookrightarrow \mathbb{E}_{\{x_n\}}\left[\frac{1}{N}\sum_{n=1}(x_n-\mu_{ML})^2\right] = \ldots = \frac{N-1}{N}\sigma^2$$

$$\hookrightarrow E\left[\frac{\cancel{N}}{N-1}\frac{1}{\cancel{N}}\sum_{n=1}^{N}(x_n-\mu_{ML})^2\right] = \dots = \cancel{\frac{N}{N-1}}\,\cancel{\frac{N-1}{N}}\,\sigma^2$$

↳ Intuition:



Exercise 10: Misclassification bound

Consider two nonnegative numbers a and b, and show that, if $a \leq b$, then $a \leq \sqrt{ab}$. Use this result to show that, if the decision regions of a two-class classification problem with classes $\mathcal{C}_1, \mathcal{C}_2$ are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\overset{\text{error}}{\cancel{\text{mistake}}}) \leq \int \{p(x, \mathcal{C}_1)p(x, \mathcal{C}_2)\}^{1/2}\, dx$$

i) Start with $a \leq b$. Since $\sqrt{\cdot}$ is monotonic,

$$\Leftrightarrow \sqrt{a} \leq \sqrt{b}$$

$$\overset{\cdot\sqrt{a}}{\Leftrightarrow} \quad a \leq \sqrt{a \cdot b}$$

ii) $p(\text{error}) = p(\hat{y} \neq y)$

$$\mathbb{1}_{\hat{y}=-1} = \begin{cases} 1 & \text{if } \hat{y}=-1 \\ 0 & \text{, else} \end{cases}$$

$$= \int p(y=1|x)\,\mathbb{1}_{\hat{y}=-1}\,p(x)\,dx$$
$$+ \int p(y=-1|x)\,\mathbb{1}_{\hat{y}=+1}\,p(x)\,dx$$

Since we're Bayes-optimal, we have that:

- if $\hat{y}=-1$ (first integral), then:

$$\overset{a}{p(y=1|x)} \leq \overset{b}{p(y=-1|x)}$$

- if $\hat{y}=1$ (second integral), then:

$$p(y=-1|x) \leq p(y=1|x)$$

↳ Using ($a \leq b \Rightarrow a \leq \sqrt{ab}$) we write:

$$p(\text{error}) \leq \int \{\overset{a}{p(y=1|x)} \cdot \overset{b}{p(y=-1|x)}\}^{1/2} \cdot \mathbb{1}_{\hat{y}=-1} \cdot p(x)\,dx$$
$$+ \int \{p(y=-1|x) \cdot p(y=1|x)\}^{1/2} \cdot \mathbb{1}_{\hat{y}=1} \cdot p(x)\,dx$$

$$+\int \{ P(Y=-1|x) \cdot P(Y=1|x) \}^{1/2} \cdot \mathbb{1}_{\hat{y}=1} \cdot p(x)\, dx$$

$$= \int \{ P(Y=1|x) \cdot P(Y=-1|x) \}^{1/2} \underbrace{\left(\mathbb{1}_{\hat{y}=1} + \mathbb{1}_{\hat{y}=-1} \right)}_{=1} p(x)\, dx$$

$$= \int \{ P(Y=1|x) \cdot p(x) \cdot P(Y=-1)|X) \cdot p(x) \}^{1/2}\, dx$$

$$= \int \{ P(Y=1, x) \cdot P(Y=-1, x) \}^{1/2}\, dx$$

Exercise 15: Decision boundary

Consider the following decision rule for a two-category one-dimensional problem: Decide $\mathcal{C}_1$ if $x > \theta$; otherwise decide $\mathcal{C}_2$.

i) Show that the probability of error for this rule is given by

$$P(error) = P(\mathcal{C}_1) \int_{-\infty}^{\theta} p(x|\mathcal{C}_1)dx + P(\mathcal{C}_2) \int_{\theta}^{\infty} p(x|\mathcal{C}_2)dx$$

ii) By differentiating, show that a necessary condition to minimize $P(error)$ is that θ satisfy

$$p(\theta|\mathcal{C}_1)P(\mathcal{C}_1) = p(\theta|\mathcal{C}_2)P(\mathcal{C}_2) \tag{8}$$

iii) Does equation 8 define θ uniquely?

iv) Give an example where a value of θ satisfying the equation actually maximizes the probability of error.

i) $P(error|x) = \begin{cases} P(C_2 \mid x), & \text{if we decide } C_1 \\ P(C_1 \mid x), & \text{if we decide } C_2 \end{cases}$

$$P(error) = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx \qquad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$= \int_{-\infty}^{\theta} P(C_1|x)\, p(x)\, dx \qquad \text{"we decide } C_2\text{"}$$

$$+ \int_{\theta}^{+\infty} P(C_2|x)\, p(x)\, dx \qquad \text{"we decide } C_1\text{"}$$

$$= P(C_1) \int_{-\infty}^{\theta} p(x|C_1)\, dx$$

$$+ P(C_2) \int_{\theta}^{+\infty} p(x|C_2)\, dx$$

ii) Derivative:

Reminder: 2. Fundamental Theorem of Calculus: $\frac{d}{dx} \int_{-\infty}^{x} f(t)\, dt = f(x)$

$$\frac{d\, p(error)}{d\theta} = P(C_1) \cdot p(\theta|C_1) - p(C_2) \cdot p(\theta|C_2) \overset{!}{=} 0$$

$\Leftrightarrow \quad P(C_1)\, p(\theta | C_1) = P(C_2) \cdot p(\theta | C_2)$

iii) No, can be true over a range of θ.

iv) If $p(x|C_1) \sim N(1,1)$ and $p(x|C_2) \sim N(-1,1)$ and $P(C_1) = P(C_2) = 1/2$ then, although $\theta = 0$ satisfies the cond., $p(\text{error})$ has a maximum at 0.