VE414 Lecture 6

Jing Liu

UM-SJTU Joint Institute

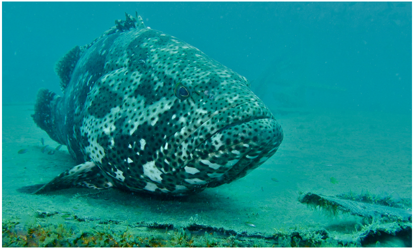
May 30, 2019

- Having various posteriors distributions is one thing,

$$f_{Y|X}; \quad f_{\{\mathbf{Y},\mathbf{Z}\}|\mathbf{X}}; \quad f_{\mathbf{Z}|\{\mathbf{X},\mathbf{Y}\}}; \quad f_{\mathbf{Z}|\mathbf{Y}}$$

but how to scientifically based our decision on them is another.

Q: How much do you know about groupers?



According to a certain chain of seafood markets, a tiger grouper is at least twice as expensive as a greasy grouper, but they are put in the same tank.

- Suppose there are only two types of fish in a very large tank

$$Y = \begin{cases} 0 & \text{if tiger,} \\ 1 & \text{if greasy.} \end{cases}$$

- In a traditional frequentist setting, knowing nothing about groupers but

$$\mathrm{P}(y = 0) > \mathrm{P}(y = 1)$$

a simple decision rule is to say it is a tiger grouper if we are forced to decide.

Q: Can you see any issue with this simple decision rule?

- Clearly it is not ideal, because that would mean assigning all fishes we catch into the same class, and the decision rule is not useful at all if we want to have a greasy grouper or if we have two distinct fishes in one catch.
- However, this is best we can do when no other information is available.
- In practice, we often have some other information available, e.g. length, the question then is how to make a better decision based on the information.

- Suppose we have the information that tiger's length has the distribution

$$X_t \sim \mathrm{Normal}\,(3, 1)$$

and greasy's length has the distribution

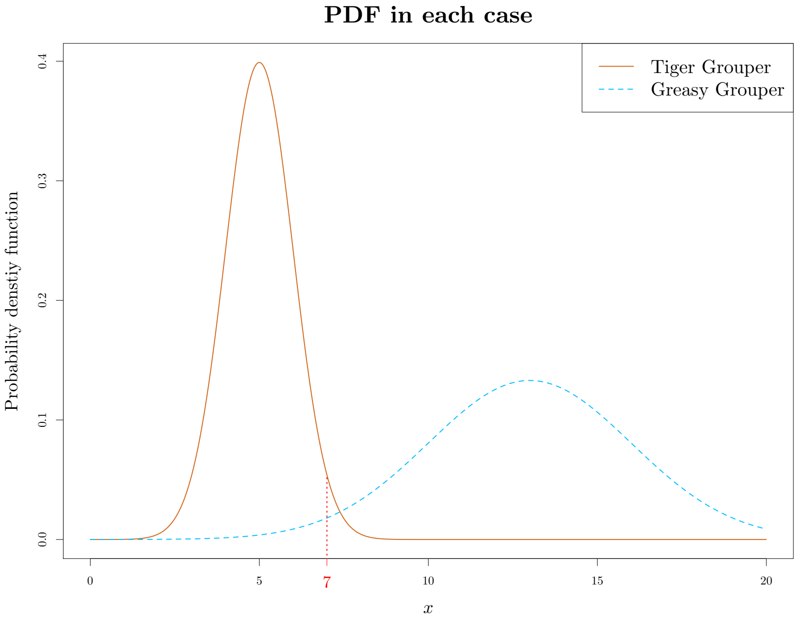
$$X_g \sim \mathrm{Normal}\,(130, 9)$$

Q: Given a particular fish is 7 units in length, what would you conclude?

- If you are forced to make a decision on what kind of fish a 7 units long fish is, it is only natural to conclude it is a tiger grouper since the probability of observing a greasy fish of a similar size extremely small.
- Now imagine we have something more realistic,

$$X_t \sim \mathrm{Normal}\,(5, 1) \qquad \text{and} \qquad X_g \sim \mathrm{Normal}\,(13, 9)$$

Q: How can we decide when we have such information in general?


PDF in each case
Tiger Grouper
Greasy Grouper
Probability denstiy function
0.0
0.1
0.2
0.3
0.4
0
5
7
10
15
20
x

- One way to understand our intuition is in the following limit sense

$$\lim_{\epsilon\to0^+}\frac{\mathrm{P}\left(7-\epsilon<X\leq7+\epsilon\mid Y=0\right)}{\mathrm{P}\left(7-\epsilon<X\leq7+\epsilon\mid Y=1\right)}=\lim_{\epsilon\to0^+}\frac{F_{X_t}\left(7+\epsilon\right)-F_{X_t}\left(7-\epsilon\right)}{F_{X_g}\left(7+\epsilon\right)-F_{X_g}\left(7-\epsilon\right)}$$
$$=\frac{f_{X_t}(7)}{f_{X_g}(7)}$$

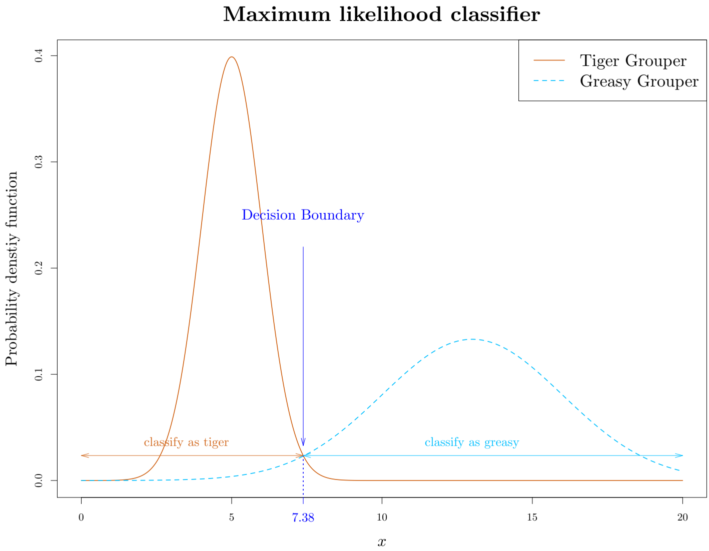
from which we see that we should choose the fish with a larger likelihood.

Q: Does it remind you of something that frequentist uses very often?

$$\begin{aligned}\mathcal{L}\left(y;x\right)&=(1-y)\cdot f_{X_t}+y\cdot f_{X_g}\qquad\text{for}\quad y\in\{0,1\}\\&=(1-y)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)+y\cdot\frac{1}{\sqrt{18\pi}}\exp\left(-\frac{(x-13)^2}{18}\right)\end{aligned}$$

- Given an observed length, say $x=7$, then y can be treated as a parameter in the above likelihood function, in other words, we are essentially basing our decision on MLE. This is known as maximum likelihood classifier.

Q: How should the decision boundary change with the prior $f_Y(y=0) = 2/3$?



Q: How would you justify your answer using Bayesian analysis?

$$f_{Y|X=x} \propto \mathcal{L}(y;x) \cdot f_Y(y)$$

- We classify the fish as a tiger grouper if

$$f_{Y|X}(0 \mid 7) > f_{Y|X}(1 \mid 7)$$

which is known as maximum a poseteriori classifier (MAP classifier).

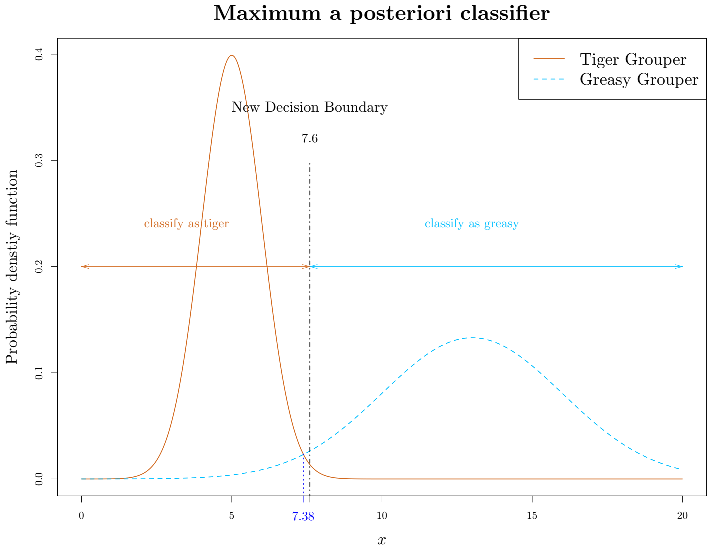
- Notice the posterior gives probabilities explicitly, thus we can solve for x

$$f_{Y=0|X=x} > f_{Y=1|X=x}$$

$$\mathcal{L}(0;x) \cdot f_Y(0) > \mathcal{L}(1;x) \cdot f_Y(1)$$

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right) \cdot \frac{2}{3} > \frac{1}{\sqrt{18\pi}} \exp\left(-\frac{(x-13)^2}{18}\right) \cdot \frac{1}{3}$$
$$\implies x < 7.6$$

Q: In what sense is the MAP classifier better than the ML classifier?



Decision as Classification

- To build a decision problem formally, we need to define the possible states:

$$y \in \mathcal{S}$$

  where y is unknown and is treated as random, so $\mathcal{S}$ is the range of Y.

- A decision rule is a function that takes an input $x \in \mathcal{D}$, i.e. observed data,

$$\delta \colon \mathcal{D} \to \mathcal{S}$$

  to a value $\delta(x)$ in the states $\mathcal{S}$.

- A loss function is a function gives the loss/cost $C\left(y, \delta(x)\right)$

$$C \colon \mathcal{S} \times \mathcal{S} \to \mathbb{R}$$

  by classifying a certain case as $\delta(x) \in \mathcal{S}$ when the true state is $y \in \mathcal{S}$.

- The risk of a decision rule δ is defined as the expected loss of using it

$$R\left[\delta\right] = \mathbb{E}\left[C\left(Y, \delta(X)\right)\right]$$

  with respect to a certain loss function C.

- Notice the expectation is over both random variables, X and Y, in general.
- We say a decision rule is optimal if

$$\delta = \arg\min_{\delta} R\left[\delta\right]$$

  which involves optimisation over a set of functions, not a set of values.

- Using the law of total expectation, we can reduce it to something simpler

$$\begin{aligned}\delta(x) &= \left\{ \arg\min_{\delta} \mathbb{E}\left[C\left(Y, \delta(X)\right)\right] \right\}\bigg|_{X=x} \\ &= \left\{ \arg\min_{\delta} \mathbb{E}\Big[\mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X\big]\Big] \right\}\bigg|_{X=x}\end{aligned}$$

- By the definition of conditional expectation, we have

$$\delta(x) = \left\{ \arg\min_{\delta} \mathbb{E}\Big[\mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X\big]\Big] \right\}\Bigg|_{X=x}$$

$$= \left\{ \arg\min_{\delta} \int_{-\infty}^{\infty} \mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X\big] \cdot f_X\left(x\right)\, dx \right\}\Bigg|_{X=x}$$

- Since $f_{Y|X} \geq 0$, the minimiser δ is the rule that minimises the integrand

$$\delta = \arg\min_{\delta} \int_{-\infty}^{\infty} \mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X\big] \cdot f_X\left(x\right)\, dx$$

$$= \arg\min_{\delta} \mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X\big] \qquad \text{for every possible } x \in \mathcal{D}.$$

- Since our state space $\mathcal{S}$ is discrete, the conditional expectation is given by

$$\mathbb{E}\big[C\left(Y, \delta(X)\right) \mid X = x\big] = \sum_{y \in \mathcal{S}} C\left(Y, \delta(x)\right) f_{Y|X}\left(y \mid x\right)$$

- Therefore, the optimal decision rule δ is the rule that minimises

$$\sum_{y \in \mathcal{S}} C\left(Y, \delta(x)\right) f_{Y \mid X}\left(y \mid x\right) \qquad \text{for every possible } x \in \mathcal{D}.$$

- Let us denote the value of the optimal decision rule by

$$z = \delta(x)$$

- We establish the function δ by defining z for each possible $x \in \mathcal{D}$, that is

$$z = \delta(x) = \arg\min_{z \in \mathcal{S}} \sum_{y \in \mathcal{S}} C\left(Y, z\right) f_{Y \mid X}\left(y \mid x\right) \qquad \text{for any } x \in \mathcal{D}.$$

- Hence the optimal decision rule evaluated at $x \in \mathcal{D}$ is given by

$$\delta(x) = \left\{ \arg\min_{\delta} \mathbb{E}\left[C\left(Y, \delta(X)\right)\right] \right\}\Big|_{X=x}$$

$$= \arg\min_{z \in \mathcal{S}} \sum_{y \in \mathcal{S}} C\left(y, z\right) f_{Y \mid X}\left(y \mid x\right)$$

- Now suppose the loss function is uniform in the following sense,

$$C(y, z) = \begin{cases} 1, & z \neq y, \\ 0, & z = y. \end{cases}$$

this loss function is known as the zero-one loss function.

- When the loss function is uniform, the optimal decision rule is

$$\begin{aligned} \delta(x) &= \arg\min_{z \in \mathcal{S}} \sum_{y \in \mathcal{S}} C\left(y, z\right) f_{Y \mid X}\left(y \mid x\right) \\ &= \arg\min_{z \in \mathcal{S}} \sum_{y \neq z} 1 \cdot f_{Y \mid X}\left(y \mid x\right) \\ &= \arg\min_{z \in \mathcal{S}} \left(1 - f_{Y \mid X}(z \mid x)\right) \\ &= \arg\max_{z \in \mathcal{S}} f_{Y \mid X}(z \mid x) \end{aligned}$$

the MAP classifier, aka maximum a posteriori (MAP) rule in decision theory.

- The probability of error is defined as

$$P_e = \mathrm{P}\left(\delta(X) \neq Y\right)$$

- In terms of our grouper example, which has only two states

$$y \in \{0, 1\}$$

corresponding to tiger grouper and greasy grouper, respectively,

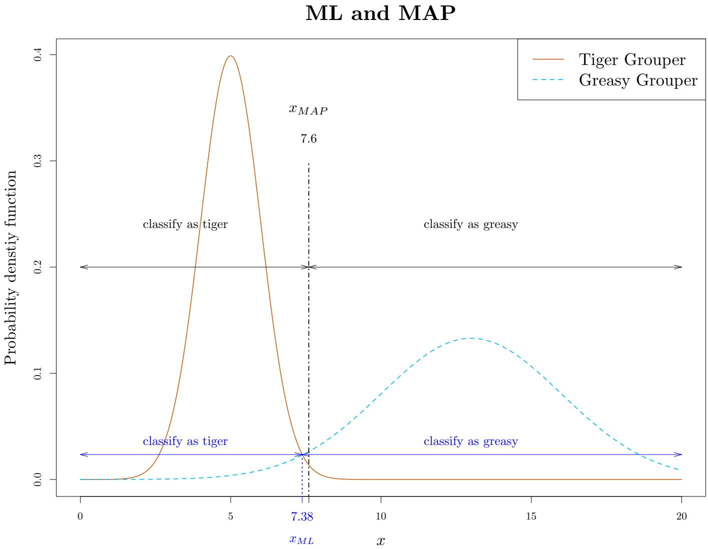
$$\text{ML}: \quad P_e = \frac{2}{3} \cdot \mathrm{P}\left(X > x_{ML} \mid Y = 0\right) + \frac{1}{3} \cdot \mathrm{P}\left(X \leq x_{ML} \mid Y = 1\right)$$

$$\text{MAP}: \quad P_e = \frac{2}{3} \cdot \mathrm{P}\left(X > x_{MAP} \mid Y = 0\right) + \frac{1}{3} \cdot \mathrm{P}\left(X \leq x_{MAP} \mid Y = 1\right)$$

where $x_{ML} = 7.38$ and $x_{MAP} = 7.6$ are the respective decision boundary.

- It can be shown that MAP is also optimal in the sense of minimising P_e.

Q: Why is the probability of error minimised by MAP?



- Recall the zero-one loss function, in which $z = \delta(x)$,

$$C(y, z) = \begin{cases} 1, & z \neq y, \\ 0, & z = y. \end{cases}$$

- The error probability is essentially equivalent to the risk under zero-one loss.

$$P_e = \mathrm{P}\left(\delta(X) \neq Y\right) = \mathbb{E}\Big[C\left(Y, \delta(X)\right)\Big] = R\left[\delta\right]$$

- Since MAP minimises the risk under zero-one loss, it also minimises P_e.

Q: Can you imagine a case where MAP is not ideal, in other words, minimising the risk under zero-one loss or the probability of error P_e is not ideal?

- For example, in terms of detecting cancer, minimising P_e is not ideal. Since false negatives are far more damaging than false positives. It is better to use

$$R = \lambda \mathrm{P}\left(\text{False Positive}\right) + (1 - \lambda)\mathrm{P}\left(\text{False Negative}\right)$$

where $\lambda \in (0, 1)$.

- The parameter $\lambda \in (0,1)$ is assigned to weight each error.

$$\begin{aligned} R &= \lambda \mathrm{P}\left(\text{False Positive}\right) + (1-\lambda)\mathrm{P}\left(\text{False Negative}\right) \\ &= \lambda \mathrm{P}(Y=0)\mathrm{P}\left(\delta(X)=1 \mid Y=0\right) \\ &\qquad + (1-\lambda)\mathrm{P}(Y=1)\mathrm{P}\left(\delta(X)=0 \mid Y=1\right) \end{aligned}$$

- In terms of the loss function, it corresponds to the following

$$C(y,z) = \begin{cases} \lambda, & y=0, z=1, \\ 1-\lambda, & y=1, z=0, \\ 0, & z=y. \end{cases}$$

- Hence the optimal decision rule is given by the solution of the following

$$\begin{aligned} z = \delta(x) &= \arg\min_{z\in\mathcal{S}} \sum_{y\in\mathcal{S}} C\left(y,z\right) f_{Y\mid X}\left(y \mid x\right) \\ &= \arg\min_{z\in\mathcal{S}} \left\{ z\lambda f_{Y\mid x}\left(0 \mid x\right) + (z-1)(1-\lambda) f_{Y\mid x}\left(1 \mid x\right) \right\} \end{aligned}$$

- Therefore, in this case, we use the following decision rule

$$\begin{array}{llll} \text{If} & \lambda f_{Y|X}(0 \mid x) > (1-\lambda) f_{Y|X}(1 \mid x), & \text{then} & z = \delta(x) = 0 \\ \text{If} & \lambda f_{Y|X}(0 \mid x) \leq (1-\lambda) f_{Y|X}(1 \mid x), & \text{then} & z = \delta(x) = 1 \end{array}$$

  which means it is different from MAP unless $\lambda = 1/2$.

- Of course, the same idea can be extended to a larger state space

$$y \in \mathcal{S} = \{0, 1, 2, \ldots, n\}$$

  or for a more complicated loss function

$$C(y, z)$$

- In practice, we will have a vector $\mathbf{x}$ of features instead of just a scalar x, e.g.

$$f_{\mathbf{X}|Y} = (2\pi)^{-k/2} \left(\det \boldsymbol{\Sigma}_y\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_y\right)^{\mathrm{T}} \boldsymbol{\Sigma}_y^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_y\right)\right)$$

- The essential idea reminds the same, our optimal decision rule is given by

$$\delta = \arg\min_{\delta} R\left[\delta\right]$$

- Under zero-one loss, MAP is optimal

$$z = \delta(\mathbf{x}) = \arg\max_{z \in \mathcal{S}} f_{Y|\mathbf{X}}(z \mid \mathbf{x})$$

- If uniform prior is used, MAP reduces to ML

$$\begin{aligned} z = \delta(\mathbf{x}) = \arg\max_{z \in \mathcal{S}} f_{Y|\mathbf{X}}(z \mid \mathbf{x}) &= \arg\max_{z \in \mathcal{S}} \frac{f_{\mathbf{X}|Y}(\mathbf{x} \mid z) f_Y(z)}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \arg\max_{z \in \mathcal{S}} f_{\mathbf{X}|Y}(\mathbf{x} \mid z) \end{aligned}$$

- If errors have different weight/nonuniform loss, then MAP is not optimal e.g.

$$z = \delta(\mathbf{x}) = \arg\min_{z \in \mathcal{S}} \left\{ z\lambda f_{Y|\mathbf{X}}\left(0 \mid \mathbf{x}\right) + (z-1)(1-\lambda) f_{Y|\mathbf{X}}\left(1 \mid \mathbf{x}\right) \right\}$$