Solutions for HW1:

Written part:

1.

(1). Predicting uniformly at random means scientists will label the fish in any one of the five labels with the same probabilities. So, for each species, scientists have 20% probability to do the right prediction. That is, the probability of incorrectly labeling a crab is:

$$25\% * 80\% + 5\% * 80\% + 20\% * 80\% + 35\% * 80\% + 15\% * 80\% = 0.8$$

(2). Scientists can predict either label 1 or 4 to improve their accuracy:
Label 1: error rate = 1- 0.25 = 0.75
Label 4: error rate = 1- 0.35 = 0.65

2.

This question can be solved easily using the function from spicy:

```
>>> from scipy.stats import norm, multivariate_normal
>>> import numpy as np
>>>
>>> x = np.array([11.1, 27.8])
>>>
>>> y_1 = multivariate_normal.pdf(x, mean=[29.42, 33.98], cov=[[50.56, 57.49], [57.49, 65.54]])
>>> y_2 = multivariate_normal.pdf(x, mean=[33.88, 37.80], cov=[[46.79, 52.19], [52.19, 58.59]])
>>>
>>> print(y_1)
2.1660889714243452e-277
>>> print(y_2)
2.1520877668415265e-141
>>>
```

As you can in this image, *multivariate_normal.pdf()* function calculate the probability using the specified means and covariance. Obviously, y_2 (orange) is larger than y_1 (blue). So, the color of the crab is most likely to be orange.

3.

(1). $P(HHTHH) = \theta * \theta * (1-\theta) * \theta * \theta = \theta^4(1-\theta)$

(2). $log\, P(HHTHH) = 4 * log\, \theta + log(1-\theta)$

(3). From the previous questions we know:

$$arg\, max_\theta\, p(HHTHH|\theta) = arg\, max_\theta\, \theta^4(1-\theta)$$
$$= arg\, max_\theta\, (4\, log\, \theta + log(1-\theta))$$

Using derivative we and make it equal to zero:

$$4/_\theta = 1/_{1-\theta}$$

So, θ should be 0.8.

Coding part:

**No marks have been awarded to those who have used libraries other than the ones mentioned in the announcement or on piazza.

The estimated μ_0 and Σ_0 of the Gaussian for the Alaskan salmon should be:

$$\begin{cases} \mu_0 = [99.22, 428.64] \\ \Sigma_0 = [[264.35, -212.54], [-212.54, 1386.23]] \end{cases}$$

The estimated μ_0 and Σ_0 of the Gaussian for the Canadian salmon should be:

$$\begin{cases} \mu_0 = [136.93, 366.64] \\ \Sigma_0 = [[338.24, 162.82], [162.82, 712.85]] \end{cases}$$

The true label of the fish in the table:
[C, A, A, C, A, C, C, C, A, A]
The predicted label of the fish in the table:
[C, A, A, C, A, C, A, C, A, A]

So, the accuracy is 90%.

For the coding part, both biased and unbiased estimates are accepted (for Σ_0).

$$\Sigma_0 \ can\ be\ [[270.36, -217.37], [-217.37, 1417.73]]\ for\ the\ Alaskan\ salmon$$

$$\Sigma_0 \ can\ be\ [[345.93, 166.52], [166.52, 729.05]]\ for\ the\ Canadian\ salmon$$

Using these slightly different Σ_0 will lead to the same result for predictions. By the way, you can still use the *multivariate_normal.pdf()* function to calculate the probabilities to decide which species fish belongs to.