Question 1 (B, T, ?)

a) We need to compute & compare the posterior probabilities for the 2 possibilities G (Greenland) or P (Poland):

$P(G|B,T)$ versus $P(P|B,T)$.

We have learned that posterior probabilities are given by Bayes rule:

$$P(G|B,T) = \frac{P(B,T|G)\cdot P(G)}{P(B,T)} \qquad (1)$$

Likewise,

$$P(P|B,T) = \frac{P(B,T|P)\cdot P(P)}{P(B,T)} \qquad (2)$$

We also know these sum to 1, so it is enough to compute either one of them and compare it to 1/2.

Or, alternatively we can compare (1) & (2) by comparing their numerators, since both have the same denominator.
Here I will take this route (as it seems slightly shorter).

We need to estimate $P(B,T|G)$, $P(G)$, $P(B,T|P)$, and $P(P)$.
The question a) asked us to use the Naive Bayes assumption.
Naive Bayes assumption is to assume that $P(B,T|G) = P(B|G)\cdot P(T|G)$
and $P(B,T|P) = P(B|P)\cdot P(T|P)$.
We plug these into the numerators of (1) & (2):

Numerator of (1) with Naive Bayes assumption:
$$P(B|G)\cdot P(T|G)\cdot P(G)$$
Numerator of (2) with Naive Bayes assumption:
$$P(B|P)\cdot P(T|P)\cdot P(P).$$
Now, we need to estimate: $P(B|G)$, $P(T|G)$, $P(G)$, $P(B|P)$, $P(T|P)$, $P(P)$.
We use the provided training set to estimate these parameters.

The estimate of $P(B|G)$ is the fraction of times we observe 'B' out of all those examples that belong to class 'G'. So,

$$P(B|G) = \frac{8}{16} = \frac{1}{2}$$

Likewise,

$$P(T|G) = \frac{10}{16} = \frac{5}{8}$$

$$P(B|P) = \frac{4}{8} = \frac{1}{2}$$

$$P(T|P) = \frac{5}{8}.$$

The estimates of the prior probabilities $P(G)$ and $P(P)$ are simply the fraction of examples that belong to the class "G" and "P" respectively:

$$P(G) = \frac{2}{3}$$

$$P(P) = \frac{1}{3}$$

Plugging these back, the numerator of (1) is:

$$\frac{8}{16} \cdot \frac{10}{16} \cdot \frac{2}{3} = 0.2083$$

Numerator of (2):

$$\frac{4}{8} \cdot \frac{5}{8} \cdot \frac{1}{3} = 0.1042.$$

Which one is larger?
The numerator of (1) is larger.
And (1) was the posterior prob. of 'G'
Therefore, the MAP answer with NB is:

'Greenland'

b) We need to compute the MAP answer <u>without</u> NB assumption.
Again, we need to compare the two posterior probabilities (1) and (2), and again this is equivalent to comparing their numerators.
But now we cannot assume anything about $P(B,T|G)$ and $P(B,T|P)$.
We need to estimate these joint conditional probabilities from the training set provided.

The estimate of $P(B,T|G)$ is the fraction of times we observe B and T together out of all those examples that belong to class 'G'.

So,
$$P(B,T|G) = \frac{2}{8}$$

Likewise,
$$P(B,T|P) = \frac{3}{8}$$

And the prior probabilities remain unchanged, so
$$P(G) = \frac{2}{3}, \quad P(P) = \frac{1}{3} \text{ as before.}$$

We plug these back into (1) and (2), and get:

Numerator of (1): $\frac{2}{8} \cdot \frac{2}{3} = \frac{4}{24}$

Numerator of (2): $\frac{3}{8} \cdot \frac{1}{3} = \frac{3}{24}$

The first of these two is larger. So, the MAP answer without NB assumption is still:

'Greenland'.

Observe, in b) we needed to estimate joint probabilities of 2 attribute values occurring together. If we had not just 2 but, say, 100 attributes, we would need to look for their occurrence together — and it is in general rare to find the exact configuration of such tuples occurring together. We would need a lot of training examples to have any chance to find them. This is why the Naive Bayes assumption is a useful approximation in such cases.

c) The ML answer is based on comparing the likelihood terms only. So we need to compare:
$$P(B,T|G) \quad \text{versus} \quad P(B,T|P)$$
The question c) asks us to use the NB assumption. Hence,
$$P(B,T|G) = P(B|G) \cdot P(T|G) = \frac{8}{16} \cdot \frac{10}{16} = \frac{1}{2} \cdot \frac{5}{8} = \frac{5}{16}$$
$$P(B,T|P) = P(B|P) \cdot P(T|P) = \frac{4}{8} \cdot \frac{5}{8} = \frac{1}{2} \cdot \frac{5}{8} = \frac{5}{16}$$
no decision can be made

d) We need to compare the likelihoods without NB:
$$P(B,T|G) = \frac{2}{8}$$
$$P(B,T|P) = \frac{3}{8}$$
the second is larger! Therefore the ML answer without NB assumption is:

"Poland"

e) We saw the answers can be different for the different methods. because:
- MAP takes into account the prior probabilities, while ML does not. So if you have reliable priors then use MAP
- the NB assumption is an approximation. So if you have reliable joint conditional probabilities then use those rather than NB. But if there is a small number of training points the estimates are not reliable.

Hence:
- in the case of a large traing set with small nr. of attributes best would be to use MAP without NB.
- in the case of a small training set with large nr of attributes it is more reliable to use ML with NB assumption.

These are the extreme cases. In intermediate cases you can use a combination e.g. MAP with NB (e.g. when you have a relatively large traing set but also a large nr. of attributes — like in text classification).

f) If we have continuous valued measurements our main tool is still to compute & compare (1) versus (2). But we can no longer estimate the likelihood term by counting! (No two people have exactly the same hair color, or hight, let alone to have both measurements identical!)

We would use a Gaussian classifier instead, by taking $p(B,T|G)$ to be a multivariate (bi-variate in this case) Gaussian with mean vector $m_G = \begin{pmatrix} m_{G1} \\ m_{G2} \end{pmatrix}$ and covariance $\underset{(2\times 2)}{\Sigma_G}$.

Likewise,
$p(B,T|P)$ would be modelled by another multivariate Gaussian with mean vector $m_P = \begin{pmatrix} m_{P1} \\ m_{P2} \end{pmatrix}$ and covariance $\underset{(2\times 2)}{\Sigma_P}$.

Then to get the estimates $p(B,T|G)$ we would compute the average (mean) color of the examples from class 'G', i.e. m_{G1} and the mean height of the examples from class 'G', i.e. m_{G2}, as well as the covariance of the examples from class 'G', which is Σ_G.

Likewise we would compute the averages and covariances of the examples in class 'P'. We would plug these in the formula of the multivariate Gaussian to get the probabilities $p(B, T \mid P)$ and $p(B, T \mid G)$.

All the rest of the calculations would be the same as those in a) – b) – c) – d). The class priors $P(G)$ and $P(P)$ are the same as there. The comparisons we need to make to get the answers are also the same.

The Naive Bayes assumption in this case would correspond to taking Σ_G and Σ_P to be diagonal. That is, we only would compute the variances of the attributes.