Question 1

Part (a). Let us expand the kernel:

$$K(x,y) = (x \cdot y + 1)^3 = (x \cdot y)^3 + 3(x \cdot y)^2 + 3(x \cdot y) + 1,$$

and by using $x \cdot y = x_1 y_1 + x_2 y_2$ (since $x, y \in \mathbb{R}^2$) we further obtain

$$\begin{aligned} K(x,y) =& (x_1 y_1 + x_2 y_2)^3 + (x_1 y_1 + x_2 y_2)^2 + (x_1 y_1 + x_2 y_2) + 1 \\ =& x_1^3 y_1^3 + 3(x_1^2 x_2)(y_1^2 y_2) + 3(x_1 x_2^2)(y_1 y_2^2) + x_2^3 y_2^3 + \\ & x_1^2 y_1^2 + 2(x_1 x_2)(y_1 y_2) + x_2^2 y_2^2 + x_1 y_1 + x_2 y_2 + 1. \end{aligned} \tag{1}$$

It is easy to see that if we introduce the following transfromation:

$$\boxed{\psi(x) = \begin{vmatrix} x_1^3 \\ \sqrt{3} x_1^2 x_2 \\ \sqrt{3} x_1 x_2^2 \\ x_2^3 \\ x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \\ x_1 \\ x_2 \end{vmatrix}}$$

then the kernel shown in Eq.(1) can be written as $K(x,y) = \psi(x) \cdot \psi(y) + 1$. Hence, the feature map ψ written above indeed represents the transformation implied by using the specified kernel $K(x,y)$.

Part (b). Full rational variety.

Part (c). In order to compute the feature map shown above, one needs to build the following quantities (*each* one will require exacly *one* multiplicatioin to compute, assuming that all the previous ones are already computed and stored): $x_1 x_2, x_1^2, x_2^2, x_2^3, x_1 x_2^2, x_1^2 x_2, x_1^3$. That's 7 multiplications to compute just one feature mapping (here we are *not* counting additional multiplications for the coefficients - let us coonsider the unscaled feature map for simplicity). Then we would need another 7 multiplications to compute the feature map for the second vector, $\mathbf{y}$. Finally, the feature space have 9 components, so in order to explicitly compute scalar product in the feature space between the vectors $\psi(x)$ and $\psi(y)$ we would need to perfrom 9 additional multiplications.

On the other hand, when computing $K(x, y)$ we just need the scalar product $x \cdot y$ in the original space $\mathbb{R}^2$, so its only 2 multiplications ($x_1 y_1$ and $x_2 y_2$); then we need to computhe the 3rd power of $(x \cdot +1)$ which can be done with 2 more multiplications.

Finally, we obtain:

$$\psi(x) \cdot \psi(y) : 7 + 7 + 9 = 23 \text{ multiplications}$$
$$K(x, y) : 2 + 2 = 4 \text{ multiplications}$$

Calculating the difference of dot product calculations, we get that

$$(\psi \text{ dimension}) - (\text{sample dimension}) = 10 - 2 = 8$$

Question 2

Lagrange function:

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(g(x, y) - 1) = 2x - y - \lambda(\frac{x^2}{4} + y^2 - 1).$$

The partial derivatives become (and they all should be equal to 0):

$$\frac{\partial \mathcal{L}}{\partial x} = 2 - \lambda x/2 = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = -1 - 2\lambda y = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\frac{x^2}{4} - y^2 + 1 = 0 \text{ (that's just the original constraint of course)}$$

and from that we immediately obtain:

$$x = \frac{4}{\lambda};\ y = -\frac{1}{2\lambda};\ \frac{x^2}{4} + y^2 = 1.$$

Next we substitute$x$, y from the first two equations into the third one to obtain:

$$\frac{16}{4\lambda^2} + \frac{1}{4\lambda^2} = 1 \ \Rightarrow\ 4\lambda^2 = 17 \ \Rightarrow\ \boxed{\lambda = \pm\frac{\sqrt{17}}{2}}.$$

The two solutions for x, y (corresponding to the two posible signs of λ) thus become:

$$x = \pm\frac{8}{\sqrt{17}}$$
$$y = \mp\frac{1}{\sqrt{17}}.$$

Substituting these two solutions to $f(x, y)$ we obtain:

$$\boxed{f(x,y) = 2x - y = \pm\frac{16}{\sqrt{17}} \pm \frac{1}{\sqrt{17}} = \pm\sqrt{17}.}$$

These are the minimum ($-\sqrt{17}$, at point $(x, y) = (-8/\sqrt{17}, 1/\sqrt{17})$) and maximum ($\sqrt{17}$, at point $(x, y) = (8/\sqrt{17}, -1/\sqrt{17})$) values of $f(x, y)$ under the specified constraint.

Question 3

Let us consider the sample S. For every point $\mathbf{x} = (x_1, x_2) \in S$ such that $c_r(\mathbf{x}) = 1$ (i.e. the point is inside the triangle described by concept c_r for a given fixed r) let us compute $r(\mathbf{x}) = \max(\mathbf{x} \cdot u, \mathbf{x} \cdot v, \mathbf{x} \cdot w)$; then let us define the agorithm that learns an estimate for r as the maximum across all such points:

$$\mathcal{A} : \hat{r} = \max_{\mathbf{x} \in S_c}(r(\mathbf{x})), \text{ where } S_c = \{\mathbf{x} : \mathbf{x} \in S \text{ and } c_r(\mathbf{x}) = 1\}.$$

Since it is guaranteed that for *any* point satisfying the concept $c_r(\mathbf{x}) = 1$ (i.e. located inside the triangle) the inequality $r(\mathbf{x}) \leq r$ must hold (by definition of the concept $c_r \in C$ given in the Problem), the same will be also true for our estimate, $\hat{r} \leq r$. That means that the trianle T_S we learn from the sample S is always embedded into the triangle $T(r)$ defined by the concept c_r with some fixed r.

In other words, the hypothesis we come up with using our algorithm does not have any false positives but can, of course, have false negatives: the points on the $\mathbb{R}^2$ plane such that $\hat{r} \leq r(\mathbf{x}) \leq r$ are inside the triangle $T(r)$ (satisfy the concept, $c_r(\mathbf{x}) = 1$) but are not classified as such by out hypothesis, so these points represent the error $h(\mathbf{x}) \neq c_r(\mathbf{x})$.

It follows from the above that all that matters for the measure of the errors is not the probability density $P(x_1, x_2)$ that describes how actual points are sampled from the plane, but only the (one-dimensional) probability density $P_r(t)$ which represents the probability (density) to sample a point $\mathbf{x}$ with $r(\mathbf{x}) = t$.

Let us fix arbitrary $\varepsilon > 0$. If for the parameter r or the concept c_r that we are trying to learn we have the measure of the whole interval $P_r(t < r) < \varepsilon$, then the error of our hypothesis $P_r(\hat{r} < t < r) < P_r(t < r) < \varepsilon$ and the error is within the requested bounds.

Hence we can assume that $P_r(t < r) > \varepsilon$. But then we can find the value r_0 such that $P_r(r_0 < t \leq r) = \varepsilon$. There are two possibilities:

1. Our estimate $\hat{r} > r_0$, but in this case $R(h) = P_r(\hat{r} < t < r) < P_r(r_0 < t < r) = \varepsilon$, and the error is again within the requested bounds already.

2. Otherwise, if $\hat{r} < r_0$, and in this case the error is $R(h) > \varepsilon$. We can compute how likely this is to happen for a sample of size m. Indeed, we need none of those m independent observations to fall into the $[r_0, r)$ interval, which has the measure ε.

This gives us $P(R(h) > \varepsilon) = (1-\varepsilon)^m < \exp(-m\varepsilon)$. For any given δ, to ensure that this probability of error does not exceed δ we must require

$$\exp(-m\varepsilon) < \delta \Rightarrow m \geq \frac{1}{\varepsilon} \ln \frac{1}{\delta}.$$

We just proved that for any given ε and δ we can ensure that the probability of the error to not exceed ε is at least $1-\delta$, $P(R(h) \leq \varepsilon) \geq 1-\delta$ if the sample size satisfies the condition given above.

This demonstrates that the proposed algorithm is polynomial.

The time complexity of the algorithm is $O(m)$ where m is a sample size. Indeed, for each point in the sample S that falls inside the triangle (i.e. $c_r(x) = 1$) we need to compute the scalar products with vectors u, v, w and then take the maximum of $r(\mathbf{x})$ across all such points. The number of $c_r(x) = 1$ points in the sample will scale linearly with m, and so will the time used by the algorithm.

Question 4

The sample error $\varepsilon = 0.2$ provides an estimate of the true error, with the 95% confidence interval $\pm 1.96\sqrt{\varepsilon(1-\varepsilon)/n} = \pm 1.96\sqrt{0.2 * 0.8/1000} \approx 0.025$.

Hence the confidence interval for the true error is $20\% \pm 2.5\%$, and the manager should report that *with 95% confidence the error is expected to be up to 22.5%*

Question 5

```
plt.legend()
plt.xscale('log')
_=plt.title('My Graph', fontdict={'size':'16'})
```