Bayesian Machine Learning

Instructor: Tim G. J. Rudner

Homework 2
Due: Friday, October 14, 11:59pm via NYU Brightspace

Model Comparison, Occam's Razor, and the Laplace Approximation
(49 marks)

The *evidence* $p(\mathcal{D}|\mathcal{M})$, also known as the *marginal likelihood*, is the probability that if we were to randomly sample parameters $\boldsymbol{\theta}$ from $\mathcal{M}$ that we would create dataset $\mathcal{D}$:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \tag{1}$$

Simple models $\mathcal{M}$ can only generate a small number of datasets, but because the marginal likelihood must normalise, it will generate these datasets with high probability. Complex models can generate a wide range of datasets, but each with typically low probability. For a given dataset, the marginal likelihood will favour a model of more appropriate complexity, as illustrated in Figure **??**.

1. (12 marks): Consider the Bayesian linear regression model,

$$y = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{z}) + \epsilon(\boldsymbol{x}) \tag{2}$$
$$\epsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma^2) \tag{3}$$
$$p(\boldsymbol{w}) = \mathcal{N}(0, \alpha^2 I) \tag{4}$$

   where the data $\mathcal{D}$ consist of N input-output pairs, $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, $\boldsymbol{w}$ is a set of linear weights which have a Gaussian prior with zero mean and covariance $\alpha^2 I$, $\boldsymbol{z}$ is a set of deterministic parameters of the basis functions $\boldsymbol{\phi}$, and σ^2 is the variance of additive Gaussian noise. Let $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ and $X = \{\boldsymbol{x}_i\}_{i=1}^N$.

   (a) (2 marks): Draw the directed graphical model corresponding to the joint distribution over all parameters.

   image.png

   (b) (2 marks): Derive an expression for the log marginal likelihood $\log p(\boldsymbol{y}|\boldsymbol{z}, X, \alpha^2, \sigma^2)$ showing all relevant steps.

   Note: I've used two approaches to try to answer the problem. The first is a bit simplier, and relies on matrixc identities to calculate the derivatives. The second uses Bishops formulas, and the derivatives are calculated with the help of a matrix calculus

calculator. To be perfectly honest, I'm not sure if they're equivalent, or if one is more correct than the other (probably the second).

Approach 1: Matrix Identities:

We can calculate the log marginal likelihood and express it as a joint probabilitiy from the graphical model:

$$\mathbb{P}\left\{y, w|z, X, \alpha^2, \sigma^2\right\} = \mathbb{P}\left\{w|\alpha^2\right\} \prod_{1=1}^{n} \mathbb{P}\left\{y|w, z, X, \sigma^2\right\}$$

Integrating with respect to the weights, we can arrive at the marginal likelihood:

$$\mathbb{P}\left\{y|z, X, \alpha^2, \sigma^2\right\} = \int \mathbb{P}\left\{w|\alpha^2\right\} \prod_{1=1}^{n} \mathbb{P}\left\{y|w, z, X, \sigma^2\right\} dw$$

In homework 1, we encountered a similar problem which involved expanding the terms, completing the square, and arriving at a new gaussian distribution. Using what we learned from HW 1, we can skip the intermediate steps and use the identity we discovered:

$$log(\mathbb{P}\left\{y|z, X, \alpha^2, \sigma^2\right\}) = log\mathcal{N}(0, \alpha^2\phi(X, z)^T\phi(X, z) + \sigma^2 I)$$

Alternatively, we can expand the definition of a multivariate gaussian to show that the following is equivalent to the above:

$$log(\mathbb{P}\left\{y|z, X, \alpha^2, \sigma^2\right\}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log(det(\Sigma)) - \frac{1}{2}y^T\Sigma^{-1}y$$

Where Σ is our covariance matrix, $\Sigma = \sigma^2\phi(X, z)\phi(X, z)^T + \alpha^2 I$

(c) (8 marks): Derive expressions for the derivatives of this log marginal likelihood with respect to *hyperparameters* $\boldsymbol{z}$, α^2, and σ^2. You can make reference to matrix derivative identities.

We have:

$$log(\mathbb{P}\left\{y|z, X, \alpha^2, \sigma^2\right\}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log(det(\Sigma)) - \frac{1}{2}y^T\Sigma^{-1}y$$

We will first direct our attention to the $-\frac{1}{2}log(det(\Sigma))$ term. We also know from matrix identities that:

$$\frac{\partial log(det(X))}{\partial \alpha} = Trace(\frac{\partial X}{\partial \alpha} \times X^{-1})$$

Calculating $\frac{\partial \Sigma}{\partial \beta}$ for $\beta \in [\alpha^2, \sigma^2, z]$ and $\Sigma = \alpha^2\phi(X, z)\phi(X, z)^T + \sigma^2 I$:

$$\begin{cases} \beta = \sigma^2 & \frac{\partial \Sigma}{\partial \beta} = I \\ \beta = \alpha^2 & \frac{\partial \Sigma}{\partial \beta} = \phi(X, z)\phi(X, z^T) \\ \beta = z & \frac{\partial \Sigma}{\partial \beta} = \alpha^2(\frac{\partial \phi}{\partial z}^T\phi + \phi^T\frac{\partial \phi}{\partial z}) \end{cases}$$

Using the above identity and computation, it follows that for the log term $(-\frac{1}{2}\frac{\partial log(det(\Sigma))}{\partial})$ the partial derivatives are as follows:

$$\begin{cases} \frac{\partial log(det(\Sigma))}{\partial \sigma^2} & -\frac{1}{2}Trace(I\Sigma^{-1}) = -\frac{1}{2}Trace(\Sigma^{-1}) \\ \frac{\partial log(det(\Sigma))}{\partial \alpha^2} & -\frac{1}{2}Trace(\phi(X,z)\phi(X,z)^T\Sigma^{-1}) \\ \frac{\partial log(det(\Sigma))}{\partial z} & -\frac{1}{2}Trace(\Sigma^{-1}\sigma^2(\frac{\partial \phi}{\partial z}^T\phi + \phi^T\frac{\partial \phi}{\partial z})) \end{cases}$$

Now we focus on the quadratic term, $-\frac{1}{2}y^T\Sigma^{-1}y$. We can leverage a nice matrix multiplication property that since Σ is symmetric, its inverse is as well. We can leverage the following: property of quadratic forms:

$$x^TMx = Trace(x^TMx) = Trace(Mxx^T) \quad \forall M \in \mathbb{R}^{n\times n},\ s.t.\ M = M^T$$

Expressing the quadratic form as trace will prove amenable during differentiation, as we now have $y^T\Sigma^{-1}y = Trace(\Sigma^{-1}yy^T)$. We will compute the derivatives using two useful properties:

$$\frac{\partial Tr(\alpha * M^{-1}yy^T)}{\partial \alpha} = Trace(M^{-1}\frac{\partial M}{\partial \alpha}M^{-1}yy^T) = -y^TM^{-1}\frac{\partial M}{\partial \alpha}M^{-1}y$$

Therefore:

$$\frac{\partial y^T(\phi(X,z)\phi(X,z)^T + I\alpha^2)y}{\partial \beta} = Trace(\Sigma^{-1}\frac{\partial \Sigma}{\partial \beta}\Sigma^{-1}yy^T)$$

And: Calculating $\frac{\partial \Sigma}{\partial \beta}$ for $\beta \in [\alpha^2, \sigma^2, z]$ and $\Sigma = \alpha^2\phi(X,z)\phi(X,z)^T + \sigma^2 I$:

$$\begin{cases} \beta = \sigma^2 & \frac{\partial \Sigma}{\partial \beta} = I \\ \beta = \alpha^2 & \frac{\partial \Sigma}{\partial \beta} = \phi(X,z)\phi(X,z)^T \\ \beta = z & \frac{\partial \Sigma}{\partial \beta} = \alpha^2(\frac{\partial \phi}{\partial z}^T\phi + \phi^T\frac{\partial \phi}{\partial z}) \end{cases}$$

Using this identity we can compute the partial derivatives for our parameters:

$$\begin{cases} -\frac{1}{2}\frac{\partial Tr(\Sigma^{-1}yy^T)}{\partial \sigma^2} & -\frac{1}{2}Trace(\Sigma^{-1}I\Sigma^{-1}yy^T) \\ -\frac{1}{2}\frac{\partial Tr(\Sigma^{-1}yy^T)}{\partial \alpha^2} & -\frac{1}{2}Trace(\Sigma^{-1}\phi(X,z)\phi(X,z)^T\Sigma^{-1}yy^T) \\ -\frac{1}{2}\frac{\partial Tr(\Sigma^{-1}yy^T)}{\partial z} & -\frac{1}{2}Trace(\Sigma^{-1}\alpha^2(\frac{\partial \phi}{\partial z}^T\phi + \phi^T\frac{\partial \phi}{\partial z})\Sigma^{-1}yy^T) \end{cases}$$

We now arrive at our result:

$$\frac{\partial log(\mathbb{P}\left\{y|z,X,\alpha^2,\sigma^2\right\})}{\partial \sigma^2} = -\frac{1}{2}Trace(\Sigma^{-1}\phi(X,z)\phi(X,z)^T\Sigma^{-1}yy^T)) - \frac{1}{2}Trace(\phi(X,z)\phi(X,z)^T\Sigma^{-1}) \tag{5}$$

$$\frac{\partial log(\mathbb{P}\left\{y|z,X,\alpha^2,\sigma^2\right\})}{\partial \alpha^2} = -\frac{1}{2}Trace(\Sigma^{-1}I\Sigma^{-1}yy^T) - \frac{1}{2}Trace(\Sigma^{-1}) \tag{6}$$

$$\begin{aligned}\frac{\partial log(\mathbb{P}\left\{y|z,X,\alpha^2,\sigma^2\right\})}{\partial z} =& -\frac{1}{2}Trace(\Sigma^{-1}\sigma^2(\frac{\partial \phi}{\partial z}^T\phi+\phi^T\frac{\partial \phi}{\partial z}\Sigma^{-1})yy^T) \\ & -\frac{1}{2}Trace(\Sigma^{-1}\sigma^2(\frac{\partial \phi}{\partial z}^T\phi+\phi^T\frac{\partial \phi}{\partial z}))\end{aligned} \tag{7}$$

Approach 2: Bishops formulas (Derivative with respect to z not calculated in this approach)

Using the Bishop Identities we can re express our function in the following form:

$$\begin{aligned}log(\mathbb{P}\left\{y|z,X,\alpha^2,\sigma^2\right\}) =& = \frac{N}{2}\log(\sigma^2)+\frac{M}{2}\log(\alpha^2) \\ & -\frac{1}{2}\log(det(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)) \\ & -\frac{\sigma^2}{2}\|y-\sigma^2\phi(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y\|_2 \\ & -\frac{\alpha^2}{2}\|\sigma^2(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}X^\top y\|_2\end{aligned}$$

Taking its gradient with respect to α^2 we have:

$$\begin{aligned}\frac{\partial}{\partial\alpha^2} =& \frac{M}{(2\alpha^2)}-\frac{(\text{Trace}((a\mathbb{I}+\sigma^2\phi^\top\phi))}{2} \\ & +\frac{((\sigma^2)^2(y^\top-\sigma^2y^\top\phi(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top)\phi(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y)}{(2\|y-\sigma^2\phi(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y\|_2)} \\ & +\frac{\|\sigma^2(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y\|_2}{2} \\ & -\frac{(\alpha(\sigma^2)^2y^\top\phi(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}(a\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y)}{(2\|\sigma^2(\alpha^2\mathbb{I}+\sigma^2\phi^\top\phi)^{-1}\phi^\top y\|_2))}\end{aligned}$$

Now with respect to σ^2:

$$
\begin{aligned}
\frac{\partial f}{\partial \sigma^2} &= \frac{N}{2\sigma^2} \\
&- \frac{(\text{Trace}(\phi^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1})}{2} \\
&+ \frac{\|y - \sigma^2 \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y\|_2}{2} \\
&- \frac{(\sigma^2 (y^\top - \sigma^2 y^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top) \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y)}{(2\|y - \sigma^2 \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y\|_2)} \\
&- \frac{((\sigma^2)^2 (y^\top - \sigma^2 y^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top) \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y)}{(2\|y - \sigma^2 \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y\|_2))} \\
&+ \frac{(\alpha^2 \sigma^2 y^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} (\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y)}{(2\|\sigma^2 (\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y\|_2)} \\
&- \frac{(\alpha^2 (\sigma^2)^2 y^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} (\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top \phi(\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y)}{(2\|\sigma^2 (\alpha^2 \mathbb{I} + \sigma^2 \phi^\top \phi)^{-1} \phi^\top y\|_2))}
\end{aligned}
$$

2. (14 marks): The posterior $p(\boldsymbol{\theta}|\mathcal{M}, \mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ will often be sharply peaked around its maximum value, as in Figure **??**. The evidence in Eq. (1) can thus be approximated by its height times its width $\sigma_{\theta|\mathcal{D}}$:

$$
\overbrace{p(\mathcal{D}|\mathcal{M})}^{\text{evidence}} \approx \overbrace{p(\mathcal{D}|\hat{\boldsymbol{\theta}}, \mathcal{M})}^{\text{data fit}} \overbrace{p(\hat{\boldsymbol{\theta}}|\mathcal{M}) \sigma_{\theta|\mathcal{D}}}^{\text{Occam factor}} \tag{8}
$$

The evidence thus naturally compartmentalizes into data fit and Occam factor terms. Suppose for simplicity that the prior is uniform on a large interval such that $p(\hat{\boldsymbol{\theta}}|\mathcal{M}) = 1/\sigma_\theta$. The Occam's factor then becomes $\frac{\sigma_{\theta|\mathcal{D}}}{\sigma_\theta}$.

(a) (2 marks): Provide an interpretation of the Occam's factor

Our expression is a way to find the model evidence by using best-fit likelihood that the model can achieve and multiplying it by an 'Occam factor'. From the reading: "The Occam factor is equal to the ratio of the posterior accessible volume of $\mathcal{M}_i$'s parameter space to the prior accessible volume, or the factor by which $\mathcal{M}_i$'s hypothesis space collapses when the data arrive."

In summary, Bayesian model comparison is a simple extension of maximum likelihood model selection: the evidence is obtained by multiplying the best-fit likelihood by the Occam factor. To evaluate the above expression, our answer will depend on several subjective assumptions: the choice probability assigned to the free parameters of each model. This is the one draw back of Bayesian statistics, "there is no such thing as inference or prediction without assumptions" – and Occam's Razor punishes models that assign a larger probability density to a greater hypothesis space.

This is exactly what happens in the figure, as the wider the prior, the larger range of models it contains, the more complexity it can capture, and less density prescribed per model on average. This is exactly how large models get penalized via low Occam factor.

(b) (4 marks): Show that if we use Laplace's method to approximate the posterior $p(\boldsymbol{\theta}|\mathcal{M}, \mathcal{D})$ as a Gaussian, then the Occam's factor becomes $p(\hat{\boldsymbol{\theta}}|\mathcal{M})\det(A)^{-1/2}$ where $A = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$.

Use this expression to interpret each of the terms in the log marginal likelihood you derived for question 1(b).

We can re-express the expression as the log of a Gaussian density, with the first and second order terms for a Tayler linear approximation:

$$p(\boldsymbol{\theta}|\mathcal{M},\mathcal{D}) \sim log(p(\hat{\boldsymbol{\theta}}|\mathcal{M},\mathcal{D})) + \nabla log(p(\hat{\boldsymbol{\theta}}|\mathcal{M},\mathcal{D}))(\theta-\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^T A(\theta-\hat{\theta})$$

Where A is the matrix of second order derivatives, AKA the hessian,

$$A = -\nabla^2 log(p(\hat{\boldsymbol{\theta}}|\mathcal{M},\mathcal{D}))$$

We can then use a couple nifty tricks: as $\hat{\theta}$ is a maxima, the first order derivative (gradient) of our expression must be 0. Encorporating this change, were left with:

$$p(\boldsymbol{\theta}|\mathcal{M},\mathcal{D}) \sim log(p(\hat{\boldsymbol{\theta}}|\mathcal{M},\mathcal{D})) - \frac{1}{2}(\theta-\hat{\theta})^T A(\theta-\hat{\theta})$$

Secondly, we have approximated the posterior as a gaussian around $\hat{\theta}$, and we know $\hat{\theta} = det(2\pi A)^{-\frac{1}{2}}$, encorporating this into our Gaussian identity we have:

$$p(\boldsymbol{\theta}|\mathcal{M},\mathcal{D}) \sim -\frac{d}{2}log(\pi) - \frac{1}{2}log(det(A)) - \frac{1}{2}(\theta-\hat{\theta})^T A(\theta-\hat{\theta})$$

We can then rewrite Bayes theorem (up to a proportionality constant) as follows:

$$\mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} = \frac{\mathbb{P}\left\{\mathcal{D},\theta|\mathcal{M}\right\}\mathbb{P}\left\{\theta|\mathcal{M}\right\}}{\mathbb{P}\left\{\theta|\mathcal{D},\mathcal{M}\right\}}$$

Substituting our identity we have found we then have:

$$\mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} = \mathbb{P}\left\{\mathcal{D},\hat{\theta}|\mathcal{M}\right\}\mathbb{P}\left\{\hat{\theta}|\mathcal{M}\right\}(2\pi^{\frac{d}{2}})det(A)^{-\frac{1}{2}}$$

If you disregard constants, the above expression asymptotic behavior as $n \to \infty$ is:

$$\mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} \xrightarrow{p} \mathbb{P}\left\{\hat{\theta}|\mathcal{M}\right\} det(A)^{-\frac{1}{2}}$$

(c) (6 marks): Derive an approximation for the log evidence $\log p(D|\mathcal{M})$ assuming a broad Gaussian prior distribution and iid observations, strictly in terms of the number of datapoints N, the number of parameters m (dimensionality of $\boldsymbol{\theta}$), and $\log p(D|\hat{\boldsymbol{\theta}})$. Show all of your work.

$$\log \mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} = \log \mathbb{P}\left\{\mathcal{D}|\hat{\theta}, \mathcal{M}\right\} + \log \mathbb{P}\left\{\hat{\theta}|\mathcal{M}\right\} + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log(det(A))$$

If take $n \to \infty$, we can observe the asymptotic behavior of the function. As the likelihood grows larger with the number of samples, (better conditioned), and the posterior is proportional to the likelihood multiplied by the prior, also grows. Dropping constants and terms not influenced by n, we have:

$$log\mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} = log\mathbb{P}\left\{\mathcal{D}|\hat{\theta}\right\} - \frac{1}{2}log(det(\nabla^2 log\mathbb{P}\left\{\mathcal{D}|\theta, \mathcal{M}\right\}))$$

As the determinant of the hessian grows exponentialy with the dimension, we can say:

$$\frac{1}{2}log(det(\nabla^2 log\mathbb{P}\left\{\mathcal{D}|\theta, \mathcal{M}\right\})) \approx N^m det(A)$$

Therefore when $n \to \infty$ and we take the log, all that's left is:

$$log\mathbb{P}\left\{\mathcal{D}|\mathcal{M}\right\} = log(\mathbb{P}\left\{\mathcal{D}|\hat{\theta}\right\} - \frac{M}{2}log(N)$$

(d) (2 marks): Relate the Hessian A to the covariance matrix of a Gaussian prior over parameters.

We know from lecture, bishop, and the past problems that the Hessian matrix, A is inversely related to the covariance matrix of a Gaussian prior over parameters, that is $A = \Sigma^{-1}$

3. (33 marks): Load the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively from `occam1.mat` and `occam2.mat` in the assignment files `a2files.zip`. Suppose we are considering three models to explain the data:

(i) $\mathcal{M}_1$: The Bayesian basis regression model of Eq.(2)-(4), but with

$$\boldsymbol{\phi}(x, \boldsymbol{z}) = \boldsymbol{\phi}(x) = (1, x, x^2, x^3, x^4, x^5)^\top \tag{9}$$

(ii) $\mathcal{M}_2$: The Bayesian basis regression model of Eq.(2)-(4), but with

$$\boldsymbol{\phi}(x, \boldsymbol{z}) = \left(\exp[-\frac{(x-1)^2}{z_1^2}], \exp[-\frac{(x-5)^2}{z_2^2}]\right)^\top, \quad \boldsymbol{z} = (z_1, z_2)^\top \tag{10}$$

(iii) $\mathcal{M}_3$: The Bayesian basis regression model of Eq.(2)-(4), but with

$$\boldsymbol{\phi}(x, \boldsymbol{z}) = \boldsymbol{\phi}(x) = (x, \cos(2x))^\top \tag{11}$$

Parts of this question involve coding. Please hand in the Matlab, Octave, or Python code you used to solve this question, along with the plots of results generated by the code used to answer the questions. This code should be succinct and include comments. Your code should not exceed 5 pages in length. Answer all questions for both $\mathcal{D}_1$ and $\mathcal{D}_2$ unless the question explicitly states otherwise.

(a) (20 marks): Using your work from question 1, write code to plot a histogram of the evidence for each of these three models, conditioned on the maximum marginal likelihood values of all the hyperparameters $\boldsymbol{z}$, α^2, and σ^2. To find these values, jointly optimize the log marginal likelihood with respect to these hyperparameters using a quasi-Newton method or non-linear conjugate gradients. Based on the histogram, which hypotheses do you believe generated the data?

Question 3 a)

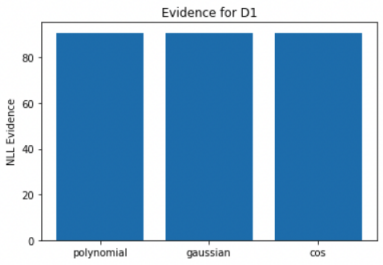
D1 Evidence

For D1, all 3 models virtually produced the same NLL values. Given the 3-way tie, we rely on the philosophy of Occam's razor that says "Nature is more likely to produce simple Models" Given this, I think the Gaussian model would be the most likely, as the polynomial model is the most complex, and gaussian distributions occur in nature all the time.

In [41]:
```
vals = []
labels = []

for key, val in small_result_map.items():
    labels.append(key)
    vals.append(val["fun"])
    print(f'Model Type: {key}, NLL Value: {val["fun"]:.5f}')
plt.title("Evidence for D1")
plt.ylabel("NLL Evidence")
plt.bar(labels,vals)
print(f"D1 Most Likely Model: {labels[vals.index(min(vals))]} NLL Value: {min(vals):.2f}")
```

```
Model Type: polynomial, NLL Value: 90.80462
Model Type: gaussian, NLL Value: 90.80462
Model Type: cos, NLL Value: 90.80462
D1 Most Likely Model: polynomial NLL Value: 90.80
```

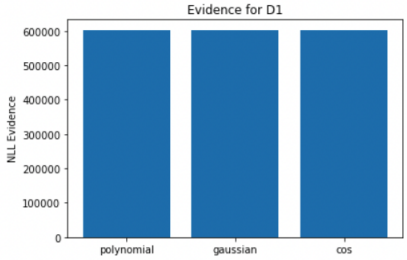


D2 Evidence

For D2, again all 3 models were close in their minimum NLL, but the model that actually became the most likely is the Cosine model.

In [42]:
```
vals = []
labels = []
#D2
for key, val in large_result_map.items():
    labels.append(key)
    vals.append(val["fun"])
    print(f'Model Type: {key}, NLL Value: {val["fun"]}')
plt.title("Evidence for D1")
plt.ylabel("NLL Evidence")
plt.bar(labels,vals)
print(f"D2 Most Likely Model: {labels[vals.index(min(vals))]} NLL Value: {min(vals):.5f}")
```

```
Model Type: polynomial, NLL Value: 602747.1446783274
Model Type: gaussian, NLL Value: 602866.2275214731
Model Type: cos, NLL Value: 602531.6640569736
D2 Most Likely Model: cos NLL Value: 602531.66406
```



(b) (8 marks): Explain why the evidence histogram might disagree with the maximum likelihood ranking of models (in general and with respect to $\mathcal{D}_1$ and $\mathcal{D}_2$).

The Bayesian approach and the MLE approach are two very different techniques that assign different levels of confidence to a model. Simply put, they will be bound to differ as a vanilla MLE approach will not have a prior, and thus, will not have a relevant Occam's factor effecting the models likelihood, aka no regularization parameter.

Take for example a situation where we have a coin and want to estimate its probability of heads as a Bernoulli R.V.. If we observed two heads in a row then, the MLE approach would give 100% probability to heads. If we used a Bayesian approach with a prior, say assuming that the coin is "fair" and that the parameter may be centered at 50%, our posterior probabilities after observing our 2 heads would effect our models likelihood. Specifically, this prior acts as a way to add some uncertainty in our model, and "regularizes" our model findings, and lowers its confidence until more data has been observed.

(c) (2 marks): Give the posterior mean and variance over the parameters of the model with highest evidence on $\mathcal{D}_2$.

Question 3 c)

In [57]:
```
z1, z2 = 0,0
mat = kernelize(x2,'cos',z1,z2)
log_alpha, log_sig,z1,z2 = large_result_map["cos"]["x"]
alpha_2 = np.exp(log_alpha)
sig_2 = np.exp(log_sig)
XTX = mat.T@mat
posterior_variance = np.linalg.inv(1/alpha_2*np.eye(XTX.shape[0])+XTX*(1/sig_2))
posterior_mean = posterior_variance@(mat.T@y2)*(1/sig_2)
print("Posterior Variance")
print(posterior_variance)

print("Posterior Mean:")
print(posterior_mean)
```

```
Posterior Variance
[[ 0.00305295 -0.00261753]
 [-0.00261753  0.19619469]]
Posterior Mean:
[0.35339692 4.79698296]
```

(d) (3 marks): What values of the *hyperparameters* maximized the marginal likelihood?

Question 3 d)

What values of the hyperparameters maximized the marginal likelihood

In [50]:
```
vals = []
labels = []

print("Hyper Parameter Order: log(Alpha^2), log(Sigma^2), z1, z2")
for key, val in small_result_map.items():
    labels.append(key)
    vals.append(val["fun"])
    print(f'D1, Model Type: {key}, Hyper-parameter values: {val["x"]}')

print("\n Hyper Parameter Order: log(Alpha^2), log(Sigma^2), z1, z2")
for key, val in large_result_map.items():
    labels.append(key)
    vals.append(val["fun"])
    print(f'D2, Model Type: {key}, Hyper-parameter values: {val["x"]}')
```

```
Hyper Parameter Order: log(Alpha^2), log(Sigma^2), z1, z2
D1, Model Type: polynomial, Hyper-parameter values: [-44.67102134   9.26940588   2.13571623   3.04473606]
D1, Model Type: gaussian, Hyper-parameter values: [0.5580811  9.2693919  0.11436398 0.05125174]
D1, Model Type: cos, Hyper-parameter values: [-27.48387196   9.26940594   2.1296608    0.63797102]

 Hyper Parameter Order: log(Alpha^2), log(Sigma^2), z1, z2
D2, Model Type: polynomial, Hyper-parameter values: [3.89480942 9.21639685 0.36278249 1.32195632]
D2, Model Type: gaussian, Hyper-parameter values: [ 2.62220424  9.2194096  -0.34688618  0.3066288 ]
D2, Model Type: cos, Hyper-parameter values: [2.5439493  9.21263112 0.5871692  1.09996998]
```

It seems that across models and data sets, the log(sigma^2) parameter converges to ~9.2

(e) Optional (1 bonus mark): Plot the posterior over each set of model weights $\boldsymbol{w}$ (using extra coding space if required) for each dataset.

Markov chain Monte Carlo (Bonus Problem)
(15 marks)

Follow Iain Murray's MCMC practical at
`http://homepages.inf.ed.ac.uk/imurray2/teaching/09mlss/handout.pdf`

Complete section 4, and answer the MCMC questions in section 5.

Hand in (1) your code for section 4, and (2) your answers to the 5 MCMC questions. The code is worth 10 marks, and the questions are worth 1 mark each. There are thus 15 marks for this part of the assignment.