Homework 0
Due: August 30

1. Suppose we have data \( D = \{(x_i, y_i)\}_{i=1}^{n} \), and \( n \) is the total number of training points. Assume we want to learn the regression model
\[
y = ax + \epsilon, \tag{1}
\]
where \( \epsilon \) is independent zero mean Gaussian noise with variance \( \sigma^2 \):
\[
\epsilon \sim \mathcal{N}(0, \sigma^2).
\]
a) Let \( y = (y_1, \ldots, y_n)^\top \) and \( X = \{x_i\}_{i=1}^{n} \). Write down the log likelihood for the whole training set, \( \log p(y|X, a, \sigma^2) \).

b) Given data \( D = \{(2, 19), (7, 62), (5, 37), (11, 94), (14, 120)\} \), find the maximum likelihood solutions for \( a \) and \( \sigma^2 \).

c) Suppose we instead consider the regression model
\[
x = by + \epsilon. \tag{2}
\]
Is the maximum likelihood solution for \( b = \frac{1}{a} \)? Explain why or why not – with derivations if necessary.

d) Returning to the model of Eq. (1), write down the multivariate pdf for \( p(y|X, a, \sigma^2) \) in terms of \( y, X, a, \sigma^2 \). (Hint: there is a short way to derive this, and a longer way involving multiplications of univariate pdfs).

e) Suppose we place a prior distribution on \( a \) such that \( p(a) = \mathcal{N}(0, \gamma^2) \). Use the sum and product rules of probability to write down the marginal likelihood of the data, \( p(y|X, \sigma^2, \gamma^2) \), conditioned only on \( y, X, \sigma^2, \gamma^2 \).

f) Without explicitly using the sum and product rules, derive \( p(y|X, \sigma^2) \), by considering the properties of Gaussian distributions and finding expectations and covariances.

g) What are the maximum marginal likelihood solutions \( \hat{\sigma}^2 \) and \( \hat{\gamma}^2 \)?

h) Derive the predictive distribution for \( p(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, D) \) for any arbitrary test point \( x_* \), where \( y_* = y(x_*) \).

i) For the dataset \( D \) in (b), give the predictive mean \( \mathbb{E}[y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, D] \) and predictive variance \( \text{var}(y_*|x_*, \hat{\sigma}^2, \hat{\gamma}^2, D) \) for \( x_* = 17 \).

j) Suppose we replace \( x \) in Eq. (1) with \( g(x, w) \), where \( g \) is a non-linear function parametrized by \( w \), and \( w \sim \mathcal{N}(0, \lambda^2) \): e.g., \( g(x, w) = \cos(wx) \). Can you write down an analytic expression for \( p(y|w, X, \sigma^2, \gamma^2) \)? How about \( p(y|X, \sigma^2, \gamma^2, \lambda^2) \)? Justify your answers.
2. Short answer questions:

a) What is the difference between local and global optima?

b) Define supervised learning, unsupervised learning, semi-supervised learning, signal, noise, over-fitting, regularization, saddle point, multimodal, cluster, cross-validation, and model complexity. Give examples wherever possible.

c) What does it mean for a matrix to be positive definite?
Is the matrix \[
\begin{bmatrix}
1 & 4 \\
9 & 7
\end{bmatrix}
\] positive definite? Is this matrix invertible? Provide brief justifications.

d) What is a conjugate prior? Give a worked example. What is conjugate to \(\sigma^2\) in Q1?

e) Define the conditioning of a matrix and give an example of an operation where conditioning is relevant.

f) If \(x \sim \mathcal{N}(0, 1)\) would you expect \(E[2x]\) to be bigger, smaller, or equal to 2\(E[x]\)? Why?

g) Using pseudo-code (or Matlab or Python), provide a numerically stable solution for \(x\) in \(Ax = b\). (\(A\) is a matrix).

h) Evaluate the Kullback-Leibler divergence between two multivariate Gaussians \(p(x) = \mathcal{N}(x; \mu, \Sigma)\) and \(q(x) = \mathcal{N}(x; m, L)\).

i) Prove the (frequently useful) Woodbury matrix inversion formula \((A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}\) (3) Hint: multiply both sides by \((A + BCD)\).

j) Consider two multivariate Gaussian vectors \(x\) and \(z\) with distributions \(p(x) = \mathcal{N}(x; \mu_x, \Sigma_x)\) and \(p(z) = \mathcal{N}(z; \mu_z, \Sigma_z)\). Given \(y = x + z\), find the marginal distribution \(p(y)\).

k) For \(p(x) = \mathcal{N}(x; \mu, \Lambda)\) and \(p(y|x) = \mathcal{N}(Ax+b, L^{-1})\), derive the marginal and conditional distributions \(p(y)\) and \(p(x|y)\).

l) Derive the maximum likelihood mean \(\mu\) and covariance matrix \(K\) of an \(n\) dimensional Gaussian distribution \(\mathcal{N}(\mu, K)\). Show all matrix and vector derivative computations.

m) Draw the contours of constant probability density for a bivariate Gaussian distribution with (i) a general covariance matrix, (ii) a diagonal covariance matrix, (iii) an identity covariance matrix. Draw the contour for when the density is equal to \(\exp(-0.5)\). Label the magnitude of the major and minor axes in terms of the eigenvalues of the covariance matrix. Label the directions of the axes with relation to the eigenvectors.

n) Show that the entropy of a multivariate distribution \(p(x)\)
\[
H[x] = - \int p(x) \log p(x) dx
\]
is maximized if \(p(x)\) is Gaussian. Hint: Use Lagrange multipliers to enforce the normalization constraints and first two moments, and then perform a variational maximization.

o) Consider the eigenvector equation for the covariance matrix \(K\) of a multivariate Gaussian distribution: \(Ku_i = \lambda_i u_i\). Show that \(K\) can be expressed as an expansion in terms of its eigenvectors: \(K = \sum_{i=1}^D \lambda_i u_i u_i^T\). Then show \(K^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} u_i u_i^T\).
3. Fun with probabilities:
   a) I have $100 and I’m gambling on a series of coin flips. For each head I win $9 and for each tail I lose $4. What’s the probability that I will run out of money? Make a plot of the probability of going broke as a function of number of flips $n$ for $n = 1, 2, \ldots, 100$, showing your work.

   b) Write down a closed form expression (binomial coefficients and factorials are okay) for the probability $p(n)$ of a shared birthday in a room of $n$ people as a function of $n$. Write down numerically stable code to compute this probability. Draw a plot of $p(n)$ for $1 < n < 100$. For what $n$ does $p(n)$ first exceed 0.7? What is the value of $p(140)$ to 15 significant figures?

   c) Suppose you have an arbitrary pdf $g(x)$. $g(x)$ has no elementary antiderivative, but you can evaluate $g(x)$ at a regular grid of points $x = x_1, \ldots, x_m$ that bounds the regions where $g(x)$ has probability mass. Devise an algorithm to sample $x_i \sim g(x)$ using only a uniform random number generator. Explain your reasoning at each step using illustrations.

   d) Suppose you can play a game with probability of winning on each attempt of $p < 0.5$. Devise a betting strategy which will guarantee you a positive expected gain. Explain why this strategy wouldn’t work in the real world.

   e) Consider a simple symmetric random walk, where the walker starts at the origin, and has an equal probability of stepping in any of the $2D$ possible directions for $D$ dimensions. For example, in a $D = 1$ random walk, the walker can go left or right. For $D = 2$, the walker can go forwards, backwards, left, and right. For $D = 3$, the walker can go up, down, forwards, backwards, left, or right. Will the walker always eventually return to the origin? Answer for $D = 1, 2, 3$. Show your reasoning.

4. General questions (not graded). Respond to these questions anonymously.

   a) How would you define machine learning? (I am not looking for a correct definition, but rather how you personally perceive machine learning, perhaps in relation to statistics, computer science, engineering, etc.).

   b) Why are you taking ORIE 6741? What are you hoping to learn in the course? What about the curriculum do you find most exciting?

   c) How did you hear about this course?

   d) Is there any topic not mentioned in the syllabus that you would like to see?

   e) Have you taken a machine learning course before? If yes, please specify, with a brief description of course level and contents.

   f) Have you done machine learning research before? Feel free to elaborate.

   g) On a scale of 1 to 4 (1: proficient, 2: comfortable, 3: rusty, 4: uneasy) how do you feel about:
      (i) Coding? (and which languages do you know?) (ii) Linear Algebra? (iii) Calculus?
      (iv) Introductory Probability (probability distributions, conditional expectation, etc.)?
      (v) Introductory Statistics (logistic regression, AR models, hypothesis testing…).

Feel free to further elaborate.