Mathematics - List of questions

**## Linear Algebra**

1. What is broadcasting in connection to Linear Algebra?
2. What are scalars, vectors, matrices, and tensors?
3. What is Hadamard product of two matrices?
4. What is an inverse matrix?
5. If inverse of a matrix exists, how to calculate it?
6. What is the determinant of a square matrix? How is it calculated (Laplace expansion)? What is the connection of determinant to eigenvalues?
7. Discuss span and linear dependence.
8. What is Ax = b? When does Ax =b has a unique solution?
9. In Ax = b, what happens when A is fat or tall?
10. When does inverse of A exist?
11. What is a norm? What is L1, L2 and L infinity norm?
12. What are the conditions a norm has to satisfy?
13. Why is squared of L2 norm preferred in ML than just L2 norm?
14. When L1 norm is preferred over L2 norm?
15. Can the number of nonzero elements in a vector be defined as L0 norm? If no, why?
16. What is Frobenius norm?
17. What is a diagonal matrix? (D\_i,j = 0 for i != 0)
18. Why is multiplication by diagonal matrix computationally cheap? How is the multiplication different for square vs. non-square diagonal matrix?
19. At what conditions does the inverse of a diagonal matrix exist? (square and all diagonal elements non-zero)
20. What is a symmetrix matrix? (same as its transpose)
21. What is a unit vector?
22. When are two vectors x and y orthogonal? (x.T \* y = 0)
23. At R^n what is the maximum possible number of orthogonal vectors with non-zero norm?
24. When are two vectors x and y orthonormal? (x.T \* y = 0 and both have unit norm)
25. What is an orthogonal matrix? Why is computationally preferred? (a square matrix whose rows are mutually orthonormal and columns are mutually orthonormal.)
26. What is eigendecomposition, eigenvectors and eigenvalues?
27. How to find eigen values of a matrix?
28. Write the eigendecomposition formula for a matrix. If the matrix is real symmetric, how will this change?
29. Is the eigendecomposition guaranteed to be unique? If not, then how do we represent it?
30. What are positive definite, negative definite, positive semi definite and negative semi definite matrices?
31. What is SVD? Why do we use it? Why not just use ED?
32. Given a matrix A, how will you calculate its SVD?
33. What are singular values, left singulars and right singulars?
34. What is the connection of SVD of A with functions of A?
35. Why are singular values always non-negative?
36. What is the Moore Penrose pseudo inverse and how to calculate it?
37. If we do Moore Penrose pseudo inverse on Ax = b, what solution is provided is A is fat? Moreover, what solution is provided if A is tall?
38. Which matrices can be decomposed by ED? (Any NxN square matrix with N linearly independent eigenvectors)
39. Which matrices can be decomposed by SVD? (Any matrix; V is either conjugate transpose or normal transpose depending on whether A is complex or real)
40. What is the trace of a matrix?
41. How to write Frobenius norm of a matrix A in terms of trace?
42. Why is trace of a multiplication of matrices invariant to cyclic permutations?
43. What is the trace of a scalar?
44. Write the frobenius norm of a matrix in terms of trace?

**## Numerical Optimization**

1. What is underflow and overflow?
2. How to tackle the problem of underflow or overflow for softmax function or log softmax function?
3. What is poor conditioning?
4. What is the condition number?
5. What are grad, div and curl?
6. What are critical or stationary points in multi-dimensions?
7. Why should you do gradient descent when you want to minimize a function?
8. What is line search?
9. What is hill climbing?
10. What is a Jacobian matrix?
11. What is curvature?
12. What is a Hessian matrix?

**## Basics of Probability and Informaion Theory**

1. Compare "Frequentist probability" vs. "Bayesian probability"?
2. What is a random variable?
3. What is a probability distribution?
4. What is a probability mass function?
5. What is a probability density function?
6. What is a joint probability distribution?
7. What are the conditions for a function to be a probability mass function?
8. What are the conditions for a function to be a probability density function?
9. What is a marginal probability? Given the joint probability function, how will you calculate it?
10. What is conditional probability? Given the joint probability function, how will you calculate it?
11. State the Chain rule of conditional probabilities.
12. What are the conditions for independence and conditional independence of two random variables?
13. What are expectation, variance and covariance?
14. Compare covariance and independence.
15. What is the covariance for a vector of random variables?
16. What is a Bernoulli distribution? Calculate the expectation and variance of a random variable that follows Bernoulli distribution?
17. What is a multinoulli distribution?
18. What is a normal distribution?
19. Why is the normal distribution a default choice for a prior over a set of real numbers?
20. What is the central limit theorem?
21. What are exponential and Laplace distribution?
22. What are Dirac distribution and Empirical distribution?
23. What is mixture of distributions?
24. Name two common examples of mixture of distributions? (Empirical and Gaussian Mixture)
25. Is Gaussian mixture model a universal approximator of densities?
26. Write the formulae for logistic and softplus function.
27. Write the formulae for Bayes rule.
28. What do you mean by measure zero and almost everywhere?
29. If two random variables are related in a deterministic way, how are the PDFs related?
30. Define self-information. What are its units?
31. What are Shannon entropy and differential entropy?
32. What is Kullback-Leibler (KL) divergence?
33. Can KL divergence be used as a distance measure?
34. Define cross-entropy.
35. What are structured probabilistic models or graphical models?
36. In the context of structured probabilistic models, what are directed and undirected models? How are they represented?
37. What are cliques in undirected structured probabilistic models?

**## Confidence interval**

1. What is population mean and sample mean?
2. What is population standard deviation and sample standard deviation?
3. Why population s.d. has N degrees of freedom while sample s.d. has N-1 degrees of freedom? In other words, why 1/N inside root for pop. s.d. and 1/(N-1) inside root for sample s.d.? (Here)
4. What is the formula for calculating the s.d. of the sample mean?
5. What is confidence interval?
6. What is standard error?

Machine Learning - List of questions

**## Learning Theory**

1. Describe bias and variance with examples.
2. What is Empirical Risk Minimization?
3. What is Union bound and Hoeffding's inequality?
4. Write the formulae for training error and generalization error. Point out the differences.
5. State the uniform convergence theorem and derive it.
6. What is sample complexity bound of uniform convergence theorem?
7. What is error bound of uniform convergence theorem?
8. What is the bias-variance trade-off theorem?
9. From the bias-variance trade-off, can you derive the bound on training set size?
10. What is the VC dimension?
11. What does the training set size depend on for a finite and infinite hypothesis set? Compare and contrast.
12. What is the VC dimension for an n-dimensional linear classifier?
13. How is the VC dimension of a SVM bounded although it is projected to an infinite dimension?
14. Considering that Empirical Risk Minimization is a NP-hard problem, how does logistic regression and SVM loss work?

**## Model and feature selection**

1. Why are model selection methods needed?
2. How do you do a trade-off between bias and variance?
3. What are the different attributes that can be selected by model selection methods?
4. Why is cross-validation required?
5. Describe different cross-validation techniques.
6. What is hold-out cross validation? What are its advantages and disadvantages?
7. What is k-fold cross validation? What are its advantages and disadvantages?
8. What is leave-one-out cross validation? What are its advantages and disadvantages?
9. Why is feature selection required?
10. Describe some feature selection methods.
11. What is forward feature selection method? What are its advantages and disadvantages?
12. What is backward feature selection method? What are its advantages and disadvantages?
13. What is filter feature selection method and describe two of them?
14. What is mutual information and KL divergence?
15. Describe KL divergence intuitively.

**## Curse of dimensionality**

1. Describe the curse of dimensionality with examples.
2. What is local constancy or smoothness prior or regularization?

**## Universal approximation of neural networks**

1. State the universal approximation theorem? What is the technique used to prove that?
2. What is a Borel measurable function?
3. Given the universal approximation theorem, why can't a MLP still reach a arbitrarily small positive error?

**## Deep Learning motivation**

1. What is the mathematical motivation of Deep Learning as opposed to standard Machine Learning techniques?
2. In standard Machine Learning vs. Deep Learning, how is the order of number of samples related to the order of regions that can be recognized in the function space?
3. What are the reasons for choosing a deep model as opposed to shallow model? (1. Number of regions O(2^k) vs O(k) where k is the number of training examples 2. # linear regions carved out in the function space depends exponentially on the depth. )
4. How Deep Learning tackles the curse of dimensionality?

**## Support Vector Machine**

1. How can the SVM optimization function be derived from the logistic regression optimization function?
2. What is a large margin classifier?
3. Why SVM is an example of a large margin classifier?
4. SVM being a large margin classifier, is it influenced by outliers? (Yes, if C is large, otherwise not)
5. What is the role of C in SVM?
6. In SVM, what is the angle between the decision boundary and theta?
7. What is the mathematical intuition of a large margin classifier?
8. What is a kernel in SVM? Why do we use kernels in SVM?
9. What is a similarity function in SVM? Why it is named so?
10. How are the landmarks initially chosen in an SVM? How many and where?
11. Can we apply the kernel trick to logistic regression? Why is it not used in practice then?
12. What is the difference between logistic regression and SVM without a kernel? (Only in implementation – one is much more efficient and has good optimization packages)
13. How does the SVM parameter C affect the bias/variance trade off? (Remember C = 1/lambda; lambda increases means variance decreases)
14. How does the SVM kernel parameter sigma^2 affect the bias/variance trade off?
15. Can any similarity function be used for SVM? (No, have to satisfy Mercer’s theorem)
16. Logistic regression vs. SVMs: When to use which one?
17. ( Let's say n and m are the number of features and training samples respectively. If n is large relative to m use log. Reg. or SVM with linear kernel, If n is small and m is intermediate, SVM with Gaussian kernel, If n is small and m is massive, Create or add more fetaures then use log. Reg. or SVM without a kernel)

**## Bayesian Machine Learning**

1. What are the differences between “Bayesian” and “Freqentist” approach for Machine Learning?
2. Compare and contrast maximum likelihood and maximum a posteriori estimation.
3. How does Bayesian methods do automatic feature selection?
4. What do you mean by Bayesian regularization?
5. When will you use Bayesian methods instead of Frequentist methods? (Small dataset, large feature set)

**## Regularization**

1. What is L1 regularization?
2. What is L2 regularization?
3. Compare L1 and L2 regularization.
4. Why does L1 regularization result in sparse models? [here](https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models)

**## Evaluation of Machine Learning systems**

1. What are accuracy, sensitivity, specificity, ROC?
2. What are precision and recall?
3. Describe t-test in the context of Machine Learning.

**## Clustering**

1. Describe the k-means algorithm.
2. What is distortion function? Is it convex or non-convex?
3. Tell me about the convergence of the distortion function.
4. Topic: EM algorithm
5. What is the Gaussian Mixture Model?
6. Describe the EM algorithm intuitively.
7. What are the two steps of the EM algorithm
8. Compare GMM vs GDA.

**## Dimensionality Reduction**

1. Why do we need dimensionality reduction techniques? (data compression, speeds up learning algorithm and visualizing data)
2. What do we need PCA and what does it do? (PCA tries to find a lower dimensional surface such the sum of the squared projection error is minimized)
3. What is the difference between logistic regression and PCA?
4. What are the two pre-processing steps that should be applied before doing PCA? (mean normalization and feature scaling)

**##Basics of Natural Language Processing**

1. What is WORD2VEC?
2. What is t-SNE? Why do we use PCA instead of t-SNE?
3. What is sampled softmax?
4. Why is it difficult to train a RNN with SGD?
5. How do you tackle the problem of exploding gradients? (By gradient clipping)
6. What is the problem of vanishing gradients? (RNN doesn't tend to remember much things from the past)
7. How do you tackle the problem of vanishing gradients? (By using LSTM)
8. Explain the memory cell of a LSTM. (LSTM allows forgetting of data and using long memory when appropriate.)
9. What type of regularization do one use in LSTM?
10. What is Beam Search?
11. How to automatically caption an image? (CNN + LSTM)

**## Miscellaneous**

1. What is the difference between loss function, cost function and objective function

Deep Learning - List of questions

**## General questions**

1. How will you implement dropout during forward and backward pass?
2. What do you do if Neural network training loss/testing loss stays constant? (ask if there could be an error in your code, going deeper, going simpler…)
3. Why do RNNs have a tendency to suffer from exploding/vanishing gradient? How to prevent this? (Talk about LSTM cell which helps the gradient from vanishing, but make sure you know why it does so. Talk about gradient clipping, and discuss whether to clip the gradient element wise, or clip the norm of the gradient.)
4. Do you know GAN, VAE, and memory augmented neural network? Can you talk about it?
5. Does using full batch means that the convergence is always better given unlimited power? (Beautiful explanation by Alex Seewald: https://www.quora.com/Is-full-batch-gradient-descent-with-unlimited-computer-power-always-better-than-mini-batch-gradient-descent)
6. What is the problem with sigmoid during backpropagation? (Very small, between 0.25 and zero.)
7. Given a black box machine learning algorithm that you can’t modify, how could you improve its error? (you can transform the input for example.)
8. How to find the best hyper parameters? (Random search, grid search, Bayesian search (and what it is?))
9. What is transfer learning?
10. Compare and contrast L1-loss vs. L2-loss and L1-regularization vs. L2-regularization.

**## Machine Learning basics**

1. Can you state Tom Mitchell's definition of learning and discuss T, P and E?
2. What can be different types of tasks encountered in Machine Learning?
3. What are supervised, unsupervised, semi-supervised, self-supervised, multi-instance learning, and reinforcement learning?
4. Loosely how can supervised learning be converted into unsupervised learning and vice-versa?
5. Consider linear regression. What are T, P and E?
6. Derive the normal equation for linear regression.
7. What do you mean by affine transformation? Discuss affine vs. linear transformation.
8. Discuss training error, test error, generalization error, overfitting, and underfitting.
9. Compare representational capacity vs. effective capacity of a model.
10. Discuss VC dimension.
11. What are nonparametric models? What is nonparametric learning?
12. What is an ideal model? What is Bayes error? What is/are the source(s) of Bayes error occur?
13. What is the no free lunch theorem in connection to Machine Learning?
14. What is regularization? Intuitively, what does regularization do during the optimization procedure? (expresses preferences to certain solutions, implicitly and explicitly)
15. What is weight decay? What is it added?
16. What is a hyperparameter? How do you choose which settings are going to be hyperparameters and which are going to be learnt? (either difficult to optimize or not appropriate to learn - learning model capacity by learning the degree of a polynomial or coefficient of the weight decay term always results in choosing the largest capacity until it overfits on the training set)
17. Why is a validation set necessary?
18. What are the different types of cross-validation? When do you use which one?
19. What are point estimation and function estimation in the context of Machine Learning? What is the relation between them?
20. What is the maximal likelihood of a parameter vector $theta$? Where does the log come from?
21. Prove that for linear regression MSE can be derived from maximal likelihood by proper assumptions.
22. Why is maximal likelihood the preferred estimator in ML? (consistency and efficiency)
23. Under what conditions do the maximal likelihood estimator guarantee consistency?
24. What is cross-entropy of loss? (trick question)

**## Optimization procedures**

1. What is the difference between an optimization problem and a Machine Learning problem?
2. How can a learning problem be converted into an optimization problem?
3. What is empirical risk minimization? Why the term empirical? Why do we rarely use it in the context of deep learning?
4. Name some typical loss functions used for regression. Compare and contrast. (L2-loss, L1-loss, and Huber loss)
5. What is the 0-1 loss function? Why can't the 0-1 loss function or classification error be used as a loss function for optimizing a deep neural network? (Non-convex, gradient is either 0 or undefined. https://davidrosenberg.github.io/ml2015/docs/3a.loss-functions.pdf)

## Parameter initialization

**## Sequence Modeling**

1. Write the equation describing a dynamical system. Can you unfold it? Now, can you use this to describe a RNN? (include hidden, input, output, etc.)
2. What determines the size of an unfolded graph?
3. What are the advantages of an unfolded graph? (arbitrary sequence length, parameter sharing, and illustrate information flow during forward and backward pass)
4. What does the output of the hidden layer of a RNN at any arbitrary time \_t\_ represent?
5. Are the output of hidden layers of RNNs lossless? If not, why?
6. RNNs are used for various tasks. From a RNNs point of view, what tasks are more demanding than others?
7. Discuss some examples of important design patterns of classical RNNs.
8. Write the equations for a classical RNN where hidden layer has recurrence. How would you define the loss in this case? What problems you might face while training it? (Discuss runtime)
9. What is backpropagation through time? (BPTT)
10. Consider a RNN that has only output to hidden layer recurrence. What are its advantages or disadvantages compared to a RNNhaving only hidden to hidden recurrence?
11. What is Teacher forcing? Compare and contrast with BPTT.
12. What is the disadvantage of using a strict teacher forcing technique? How to solve this?
13. Explain the vanishing/exploding gradient phenomenon for recurrent neural networks. (use scalar and vector input scenarios)
14. Why don't we see the vanishing/exploding gradient phenomenon in feedforward networks? (weights are different in different layers - Random block intialization paper)
15. What is the key difference in architecture of LSTMs/GRUs compared to traditional RNNs? (Additive update instead of multiplicative)
16. What is the difference between LSTM and GRU?
17. Explain Gradient Clipping.
18. Adam and RMSProp adjust the size of gradients based on previously seen gradients. Do they inherently perform gradient clipping? If no, why?
19. Discuss RNNs in the context of Bayesian Machine Learning.
20. Can we do Batch Normalization in RNNs? If not, what is the alternative? (BNorm would need future data; Layer Norm)
21. **## Autoencoders**
22. What is an Autoencoder? What does it "auto-encode"?
23. What were Autoencoders traditionally used for? Why there has been a resurgence of Autoencoders for generative modeling?
24. What is recirculation?
25. What loss functions are used for Autoencoders?
26. What is a linear autoencoder? Can it be optimal (lowest training reconstruction error)? If yes, under what conditions?
27. What is the difference between Autoencoders and PCA (can also be used for reconstruction - https://stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com).
28. What is the impact of the size of the hidden layer in Autoencoders?
29. What is an undercomplete Autoencoder? Why is it typically used for?
30. What is a linear Autoencoder? Discuss it's equivalence with PCA. (only valid for undercomplete) Which one is better in reconstruction?
31. What problems might a nonlinear undercomplete Autoencoder face?
32. What are overcomplete Autoencoders? What problems might they face? Does the scenario change for linear overcomplete autoencoders? (identity function)
33. Discuss the importance of regularization in the context of Autoencoders.
34. Why does generative autoencoders not require regularization?
35. What are sparse autoencoders?
36. What is a denoising autoencoder? What are its advantages? How does it solve the overcomplete problem?
37. What is score matching? Discuss it's connections to DAEs.
38. Are there any connections between Autoencoders and RBMs?
39. What is manifold learning? How are denoising and contractive autoencoders equipped to do manifold learning?
40. What is a contractive autoencoder? Discuss its advantages. How does it solve the overcomplete problem?
41. Why is a contractive autoencoder named so? (intuitive and mathematical)
42. What are the practical issues with CAEs? How to tackle them?
43. What is a stacked autoencoder? What is a deep autoencoder? Compare and contrast.
44. Compare the reconstruction quality of a deep autoencoder vs. PCA.
45. What is predictive sparse decomposition?
46. Discuss some applications of Autoencoders.

**## Representation Learning**

1. What is representation learning? Why is it useful? (for a particular architecture, for other tasks, etc.)
2. What is the relation between Representation Learning and Deep Learning?
3. What is one-shot and zero-shot learning (Google's NMT)? Give examples.
4. What trade offs does representation learning have to consider?
5. What is greedy layer-wise unsupervised pretraining (GLUP)? Why greedy? Why layer-wise? Why unsupervised? Why pretraining?
6. What were/are the purposes of the above technique? (deep learning problem and initialization)
7. Why does unsupervised pretraining work?
8. When does unsupervised training work? Under which circumstances?
9. Why might unsupervised pretraining act as a regularizer?
10. What is the disadvantage of unsupervised pretraining compared to other forms of unsupervised learning?
11. How do you control the regularizing effect of unsupervised pretraining?
12. How to select the hyperparameters of each stage of GLUP?

**## Monte Carlo Methods**

1. What are deterministic algorithms? (nothing random)
2. What are Las vegas algorithms? (exact or no solution, random resources)
3. What are deterministic approximate algorithms? (solution is not exact but the error is known)
4. What are Monte Carlo algorithms? (approximate solution with random error)

**## Adversarial Networks**

1. Discuss state-of-the-art attack and defense techniques for adversarial models.