21. What is the singular value decomposition of a matrix A.

Singular value decomposition of matrix A is to decompose A to be a product of tranpose of one orthogonal matrix U, one diagonal matrix Σ whose diagonal entry is singular value of A and another othogonal matrix V.

$$A = U^T \Sigma V$$

22. Calculate the gradient of the following function

$$f(X) = tr\{X^T C X\}$$

Solutions starting from definition of derivative:

$$\lim_{n\to 0} \frac{f(X+n\xi) - f(X)}{n} \tag{1}$$
$$= \lim_{n\to 0} \frac{tr\{(X+n\xi)^T C(X+n\xi)\} - tr\{X^T CX\}}{n} \tag{2}$$
$$= \lim_{n\to 0} \frac{tr\{(X+n\xi)^T C(X+n\xi) - X^T CX\}}{n} \tag{3}$$
$$= \lim_{n\to 0} \frac{tr\{X^T CX + nX^T C\xi + n\xi^T CX + n^2\xi^T C\xi - X^T CX\}}{n} \tag{4}$$
$$= \lim_{n\to 0} \frac{tr\{nX^T C\xi + n\xi^T CX + n^2\xi^T C\xi\}}{n} \tag{5}$$
$$= \lim_{n\to 0} tr\{\frac{nX^T C\xi + n\xi^T CX + n^2\xi^T C\xi}{n}\} \tag{6}$$
$$= \lim_{n\to 0} tr\{X^T C\xi + \xi^T CX + n\xi^T C\xi\} \tag{7}$$
$$= tr\{X^T C\xi + \xi^T CX\} \tag{8}$$
$$= tr\{X^T C\xi + X^T C^T \xi\} \tag{9}$$
$$= tr\{X^T (C + C^T)\xi\} \tag{10}$$

23. Explain the difference of frequentist's approach and Bayesian approach.

Aspects: **objective, sequential learning, compromise**

Frequetist's approach is to use maixmum likelihood estimation to do point estimation of parameter w. Does not require preknowledge but need a large amount of data. The sequential learning is based on the sequential update formula

$$w^{(\tau+1)} = w^{(\tau)} - \xi \bigtriangledown E(w)$$

Whereas the bayesian approach is to derive a probability distribution over the parameter based on the Bayes Theorem. This approach does not necesarily require a large amount of data if the prior knowledge is provided. And it is inherently a good framework for sequential learning. The posterior in current iteration works as the prior in the next iteration.

$$p(w^{(\tau+1)}|t) = \frac{p(t|w^{(\tau)})p(w^{(\tau)})}{p(t)}$$

According to the Bayes Theorem, there is a **compromise** between the observation of dataset and the preknowledge. If the amount of data is very large or preknowledge (prior) provides no disparity for various classes, the preknowledge can be ignored. However, when the data is scarce, the preknowledge plays an important role.

24. What is S-fold cross validation? Why we use it?

A method to compare the performance of each model. Randomly group all labeled data into S sets, and run experiment for S times. Each time leave one set as testing set and others as training set.

Extreme: leave-one-out cross validation, S = N, each set contains only one piece of labeled data.

Functionality: Model Comparison. Such model can be of different form or of the same form with different parameter (such as to determine K in K-Nearest Neighbour, or determine).

25. Derive a least squre solution for linear regression model (with regularisation).

The error function for least square error with regularisation:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n) - t_n)^2 + \frac{1}{2}w^T w \tag{11}$$

$$= \frac{1}{2}(w^T) \tag{12}$$

$$w = (\lambda I - \Phi^T\Phi)^{-1}\Phi t$$

26. What is stochastic gradient descent? How to apply it to linear regression problem specifically with the sum-of-squares error function?

Stochastic gradient descent is one optimisation algorithm, which consider only the error of one piece of observed data at each iteration of training. That is where it is different from the batch gradient descent, which takes the averaged errors of all the observed data in one training iteration.

$$w^{(\tau+1)} = w^{(\tau)} - \xi \bigtriangledown E(w)$$

In sum-of-square function, we have error function for one observed data as follows,

$$E(w) = (w^T\Phi(x_n) - t_n)^2$$

whose gradient is

$$\bigtriangledown E(w) = (w^T\Phi(x_n) - t_n)\Phi(x_n)$$

Take this gradient of error function back, we have

$$w^{(\tau+1)} = w^{(\tau)} - \xi(w^T\Phi(x_n) - t_n)\Phi(x_n)$$

27. What is the bias-variance decomposition? Demonstrate the bias-variance decomposition where the error is measured by the mean squared error. What you can deduce from the result?

add some thing here..

28. What is a conjugate prior? Why we normally use conjugate prior? What is the conjugate prior for Gaussian distribution?

A prior is conjugate to the likilihood function, if the posterior distribution derived from the Bayes Theorem shares the same form (perhaps with different parameters) with prior distribution.

Use conjugate prior to simplify the sequential learning process in the sense that we make the update formula the same at all iterations. Conjugate prior of gaussian distribution is gaussian distribution.

29. What are the limitations of linear basis function models?

We have to fix the basis functions before the data is observed. It is very likely that the basis function we choose cannot adapt to the target function. And thus, the lower bound of the error may be quite large.

30. What is the curse of dimensionality? Why it can be a problem?

Curse of Dimensionality: the number of basis functions (amount of the feature mapping) grows exponentially with regard to the dimensionality D.

why it can be a problem?: the number of parameter in vector w we need to deal with is increased exponentially.

By the way, in rejection sampling, there is another curse of dimensionality: the probability of rejection increases exponentially with regard to the dimensionality of the sampled space.

31. What are the three models for decision problems? How they are different?

Linear Discriminant. Non-probalistic model. Use discriminant function, directly map the input pattern x into class decision. (e.g. FDA, Perceptron algorithm.)

Discrminative Model. Probalistic model which directly model the posterior distribution $p(C_1|x)$. And make decision based on that posterior distribution. (e.g. logistic regression)

Generative Model. Probalistic model which first model the class-conditional probability distribution $p(x|C_k)$ and prior $p(C_k)$. And finally derive the posterior distribution $p(C_k|x)$ based on Bayes theorem. (e.g. naive bayes for discrete input with conditional independence assumption.)

Generally, the generative model is the most expensive model because of the large number of parameter in class-conditional distribution $p(x|C_k)$ to be fitted.

32. What are the deficiencies of the least squares approach in linear classification? (refer Textbook P185-186)

(1). Least-squares solutions lack robustness to outliers.

(2). In some multi-class classification problems that linear decision boundaries can give excellent separation between the classes, the least squares approach will misclassify most of the points from some classes.

33. What is the idea of Fisher's Linear Discriminant? Derive the fisher's linear discriminant that projects $x \in \mathbb{R}^D$ to $x \in \mathbb{R}^{D'}$ where $D > D'$.

ADD something here...

34. What is the perceptron algorithm? Describe it in detail.

Basic idea of perceptron algorithm is to minimise the number of misclassified patterns in two-class classfication problem.

Generalised linear model:

$$y(x) = f(w^T\Phi(x)) = +1\ (w^T\Phi(x) > 0), -1\ (w^T\Phi(x) < 0)$$

Use Particular coding scheme: $+1$ for C_1, -1 for C_2.

Error function is defined:

$$E(w) = -\sum_{n\in\mathbb{M}} w^T\Phi(x_n)t_n$$

Intuitively, correctly classified pattern will not contribute to the update, while misclassified pattern do. However, after the update, the correctly classified pattern in the past may turn to be wrong.

This algorithm guaranteed to converge, applied with stochastic gradient descent:

$$-w^{(\tau+1)}\Phi(x_n)t_n = -w^{(\tau)}\Phi(x_n)t_n - (\Phi(x_n)t_n)^T\Phi(x_n)t_n < -w^{(\tau)}\Phi(x_n)t_n$$

35. What is the probalistic generative model? Describe it in detail.

We have Naive Bayes for the generative model with discrete input. The Naive Bayes make a class conditional independence assumption: all features are independent on each other conditioned on the class.

36. What is logistic regression? Describe it in detail.

Logistic Regression, in essence, is a discrminative model to solve the classfication problem. It directly model the posterior distribution $y_n = p(C_1|x_n)$. Its likelihood function is

$$p(t, X) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

37. What is the feature mapping in classfication problem? Why we need this feature mapping sometimes?

Map input pattern from input space to feature space, on which the decision evaluation is based.

Sometimes, the patterns in input space is not linearly separable but in feature space it is linearly separable. Note that the overlapping problem cannot be solved by the mapping into feature space.

38. What is Laplace Approximation? Why we need to do Laplace Approximation sometimes? Describe in details.

- Approximate p(**z**) for $z \in \mathbb{R}^M$

$$p(\mathbf{z}) = \frac{1}{Z}f(\mathbf{z}).$$

- we get the Taylor expansion

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

- where the Hessian **A** is defined as

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})\ |_{\mathbf{z}=\mathbf{z}_0}.$$

- The Laplace approximation of $p(\mathbf{z})$ is then

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}\,|\,\mathbf{z}_0, \mathbf{A}^{-1})$$

Figure 1: Laplace Approximation in vector space.

39. Describe Baysian Logistic Regression in detail.
ADD something here...

61. Show that Gaussian distribution belongs to exponential family.

- The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form
$$p(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}$$
where $\mathbf{x}$ may be scalar or vector, and may be discrete or continuous.
- Natural parameter $\boldsymbol{\eta}$
- And $\mathbf{u}$ is some function of $\mathbf{x}$.
- The function $g(\boldsymbol{\eta})$ can be interpreted as the coefficient ensuring normalisation
$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\} \, d\mathbf{x} = 1$$
- Other form with $g(\boldsymbol{\eta}) = \exp\{-G(\boldsymbol{\eta})\}$ we get
$$p(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) - G(\boldsymbol{\eta})\}$$

Normal distribution with mean μ and standard deviation σ

$$\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$
$$\boldsymbol{\eta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)^T$$
$$h(x) = \frac{1}{\sqrt{2\pi}}$$
$$\mathbf{u}(x) = (x, x^2)^T$$
$$g(\boldsymbol{\eta}) = \sqrt{-2\eta_2} \exp\left(\frac{\eta_1^2}{4\eta_2} \right)$$
$$G(\boldsymbol{\eta}) = -\frac{1}{2} \ln(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}$$

Figure 2: Demonstrate Gaussian distribution belongs to exponential family.

62. How do we sample from standard distribution?
63. Describe the method you would use to sample from the following distribution:

$$p(x = i) = \frac{1}{i} - \frac{1}{i+1}, \; i = 1, 2, 3, 4...$$

64. Describe rejection sampling in detail.
65. Describe importance sampling in detail.
66. Describe Metropolis-Hasting algorithm in detail.
67. What is the idea behind principle component analysis?

(1) PCA is an algorithm for unsupervised learning. (differ from FDA)

(2) Functionality: Dimensionality Reduction. Project data into a subspace where data objects span the most widely on the bases of subspace.

(3) Functionality: Data Decorelation. Transform the coordinate of original space and apply the scaling on each coordinate axis, such that in new coordinate system, the mean of data objects is zero and covariance matrix is identity.

68. What is a non-stationary distribution?

Non-stationary distribution is the

69. Use a Bayesian Network to represent a second-order Markov chain.

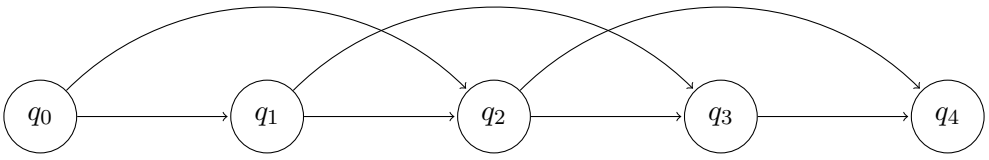

Figure 3: Bayesian Network to represent second-order Markov Chain

70. What is the difference between principle component analysis and Fisher's discriminant?

- Essence: FDA is supervised learning. PCA is unsupervised learning. (differ in existence of labeled data)
- Goal: FDA is to fulfill classification problem. PCA is to do dimensionality reduction or data decorrelation.
- Basic Idea: FDA, given the labels, maximise the between-class covariance and minimise the within-class covariance. PCA consider the maximally reserved data variance (information).

71. Prove that independence of two random variable implies uncorrelatedness. Give counter example to show that uncorrelatedness does not imply independence.

Independence: $P(x, y) = p(x)p(y)$. Uncorrelatedness: $cov(x, y) = 0$

Proof of *independence* $\rightarrow$ *uncorrelatedness*:

$$
\begin{aligned}
Cov(X, Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(x))(y - E(y))p(x, y)dxdy && (13)\\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(x))(y - E(y))p(x)p(y)dxdy && (14)\\
&= \big(\int_{-\infty}^{+\infty} (x - E(x))p(x)dx \big) \cdot \big(\int_{-\infty}^{+\infty} (y - E(y))p(y)dy \big) && (15)\\
&= E(x - E(x)) \cdot E(y - E(y)) && (16)\\
&= 0 && (17)
\end{aligned}
$$

Counter-example:

Suppose X is a normally-distributed random variable with zero mean. Let Y = X^2. Clearly X and Y are not independent. However, The covariance of X and Y is

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X^3) - 0 * E(Y) = E(X^3) = 0$$

72. Assuming K different states for each variable x, how many parameters does M-order Markov chain have?

$$K^M(K - 1)$$

73. What is a homogeneous hidden Markov Model?

Homogeneousness in hidden markov model means that all the transition probabilities are the same.

The probability distribution of latent variable is **invariant or stationary** if the **Detailed balance** is satisfied. (also called **reversible**.)

$$p^*(z)T(z, z') = p^*(z')T(z', z)$$

74. Describe Viterbi algorithm in detail.

75. What are the motivations for combining models?

- Motivation: Coming up with a very precise prediction rule can be very hard. It may be easier to come up with a number of not so precise prediction rules.
- Basic Idea: Combine models with different rules to make better prediction.