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:

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

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:

$$
\begin{aligned}
E(w) &= \frac{1}{2}\sum_{n=1}^{N}(y(x_n) - t_n)^2 + \frac{1}{2}w^T w && (11)\\
&= \frac{1}{2}(w^T) && (12)
\end{aligned}
$$

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

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?

TBD: It may not produce good result.

(refer to the lecture notes 07_Linear_Classification_1 P21-22).

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 $D > D'$.

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

Basic idea of perceptron algorithm is to minimise the number of misclassified patterns.

Generalised linear model:

$$y(x) = f(w^T \Phi(x))$$

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.

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

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.

Laplace Approximation is to derive a gaussian distribution to approximate given distribution such that two distribution has the same mode. Given arbitrary probability distribution $q(x)$ which is not gaussian:

$$q(x) = \frac{1}{Z} f(x)$$

Ignore the normalising factor Z, only focus on the $f(x)$ dependent on variable x.

Take logarithm for $f(x)$ and apply taylor expansion on it to second order at point z_0, where

$$ln(f(z)) = ln(f(z_0)) + \frac{A}{2}(z - z_0)^2$$

where $A = \bigtriangledown \bigtriangledown ln(f(z))|_{z=z_0}$ and $\bigtriangledown ln(f(z_0)) = 0$.

By exponenting both side, we derive the desired distribution which approximates the $q(x)$.

$$g(x) = (\frac{A}{2\pi})^{\frac{1}{2}} exp(\frac{A}{2}(z - z_0)^2)$$

39. Describe Baysian Logistic Regression in detail.

ADD something here...