根据**yb**老师上课讲的（听录音**+**几位大佬的笔记）（**50%**），去年学长的答案（**30%**），还有我查资料自己瞎写的（**20%**），使用者请谨慎。。如有错误，概不负责；如有挂科，概不负责。

1. Because learning problem is learning from data to get a general model which means it needs to learn general rules from few data.(not sure)
2. (胡扯)

   (1)If not locally smooth, we can not apply differential or gradient to value the model. And we must follow the gradient to tune the parameters, which is the only approach. If there is no such gradient.

   (2)If not, the problem will become exhaustive.
3. Generalizability shows whether our model can achieve good result in new in test dataset and is robust in solving real world problems.

   (1)From smoothness point of view: the model/process can encode/decode many inputs and outputs readily.

   (2)From capacity point of view: it works on not only the given data but also all possible x&y(new data, test data)
4. Soft margin: overlapping class Kernel method: mapping nonlinear relationship to linear one Regularization: dimension reduction to reduce complexity Cross validation: finding a model that generalizes well to testing data
5. Too linear/ Local solution(not feasible)

   Curse of dimensionality

   Singularity

   Difficult to project real (x,y,z) to Euclidean space
6. Philosophical: Truth cannot be proved

   Mathematical: Fermat's Last Theorem

   Physical: Three-Body problem

   Computational: Halting problem

   Numerical: Initial value sensitivity problem
7. (1)In real world, the features are heterogeneous and arbitrary. We need to find a way to measure them so that we can optimize the problem.

   (2)The data in real world are heterogeneous, complex and dirty. Most of them are not independent and not measurable. We can use some methods to deal with the dimension space and apply feature selection to make them more measurable for Euclidean space but these methods are not general enough.
8. (这块乱七八糟讲了一堆)Orthogonal/linear?(提了这个东西但是想不通)/normalizable/measurable/eigenvalue projection?/Equally expandable/Transfer a dirty set to an idealistic problem
9. (也是超级乱)feature dependence/feature selection -> graphic model(describe them as a structure)

   Markov? (总之一直在说feature point of view)
10. (这题觉得讲的答非所问)(1)Many features are redundant.

(2)Human are limited that cannot easily perceive or deal with high-dimensional problem.

(3)Many features can cancel each other which makes some of them outstanding.

11. (1) Tend to capture some important features and details.

    (2)Overfitting is worse than underfitting. Low complexity is less possible to overfit. Increase the robustness.

    (3)Ease the pain from feature selection and more easy to learn.

    (4)Less possible to get a local optimal solution and singularity.

12. (1)Loss function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. In machine learning loss function is a function that measures the difference between the prediction result and the true result.

    (2)a. Least Square: A quadratic surface b. Logistic: Sigmoid If "Logistic" means Logistic loss, then for an instance x and its label y and hypothesis h(x):

    $E(x) = \frac{1}{ln2} ln(1 + e^{-yh(x)})$

    $ylogh(x) + (1 - y)log[1 - h(x)]$

    c. Hinge: $max(0, 1 - z)$,$max(0, 1 - t \cdot y)$

    (3)a. Least Square: It makes the assumption that the data distribution is Gaussian distribution and uses a homogeneous way to view data. b. Logistic: More robust than Least Square. No assumption about data distribution. c. Hinge: More robust. No universal assumption about distance...(YB's points of view)

13. a. Using soft-margin b. Using L1 or L2 norm to reduce dimension c. Using kernel method

14. (1)Suppose the true target given x is $t(x)$, and our model is $h(x)$, the probability density of x is $p(x)$. We use the square loss, then:

    The variance:

    $E(L) = \int (h(x) - t(x))^2 p(x) dx$

    $= \int (h(x) - E[t(x)] + E[t(x)] - t(x))^2 p(x) dx$

    $= \int (h(x) - E[t(x)])^2 p(x) dx + \int (E[t(x)] - t(x))^2 p(x) dx$

    The square of bias:

    $\int (h(x) - E[t(x)])^2 p(x) dx = \int (h(x) - E[h(x)] + E[h(x)] - E[t(x)])^2 p(x) dx$

    $= \int (h(x) - E[h(x)])^2 p(x) dx + \int (E[h(x)] - E[t(x)])^2 p(x) dx$

    (2) We can use cross-validation or leave one alone to select and validate the model.

15. (a) Using regularization term. (b) Using L1 and L2 norm to reduce dimension.

16. $y = f(x) + e$ and $e \sim N(0, I)$ Thus $y \sim N(f(x_i), I)$

    The likelihood:

    $L = \prod_i p(y_i|x_i) = \prod_i \frac{1}{(2\pi)^{D/2}} exp(-\frac{1}{2}(y_i - f(x_i))^T (y_i - f(x_i)))$

    Where D is the dimension of x.

    Take log:

    $logL = -\frac{1}{2} \sum_i (y_i - f(x_i))^T (y_i - f(x_i)) + C$

    Where C is a constant. Max this equation equals to min the minus one .

    So,

$min\frac{1}{2}\sum_i(y_i-f(x_i))^T(y_i-f(x_i))$

This is the least square method.

17. Convexity means the global optimum is unique and we can use Gradient-based method easily to nd it.
18. The value of hyper-parameters like learning rate (step size).

    How to jump out of the local minimum. The convexity of the problem. Parallel computation and speed up.
19. Existence shows whether our model can converge.

    Uniqueness shows the difficulty of training.

    If the problem is convex, we can solve it easily and the global minimum always exists and is unique.

    Complexity shows the cost of training.

    Generalizability shows whether our model can achieve good result in new in test dataset and is robust in solving real world problems.
20. Logistic regression learns the posterior distribution $p(y|x)$ directly, is a discriminative method. The MLE-based generative method makes the assumption that the class density $p(x|C)$ is a Gaussian distribution and all these density shares a same covariance matrix. Then using MLE to get the parameters. The Logistic regression and the MLE-based generative method have similar formula:

    $p(y=1|x)=\frac{1}{1+exp(-\theta^Tx)}$

    Logistic regression without regularization term = Gaussian Naive Bayes
21. Suppose the prior probability of Gaussian Mixture model $P(A)=p, P(B)=1-p$ and the input $X|A\sim N(\mu_a,\sigma_a^2), X|B\sim N(\mu_b,\sigma_b^2)$

    Given an input x,

    $$P(x \text{ from } A)=\frac{P(A)P(x\sim X_A)}{P(x)}=\frac{p\frac{1}{\sqrt{2\pi}\sigma_a}e^{-\frac{(x-\mu_a)^2}{2\sigma_a^2}}}{p\frac{1}{\sqrt{2\pi}\sigma_a}e^{-\frac{(x-\mu_a)^2}{2\sigma_a^2}}+(1-p)\frac{1}{\sqrt{2\pi}\sigma_b}e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}}$$

    Let the feature vector of input data be $x=(1,x,x^2)^T$, the posterior probability of classication has the form

    $P(x \text{ from } A)=\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$

    where w is the weight vector:

    $$\mathbf{w}=\left(\ln\frac{p\sqrt{\sigma_b}}{(1-p)\sqrt{\sigma_a}}+\frac{\mu_b^2}{2\sigma_b^2}-\frac{\mu_a^2}{2\sigma_a^2},-\frac{\mu_b}{\sigma_b^2}+\frac{\mu_a}{\sigma_a^2},\frac{1}{2\sigma_b^2}-\frac{1}{2\sigma_a^2}\right)^T$$
22. (第一段是我自己扯的，因为yb讲的跟问题要求不太符合，可以参考拳师的ML的note)Logistic regression is to analyze the relationship between the probability of taking a certain value of the dependent variable and the independent variable, while linear regression is to directly analyze the relationship between the dependent variable and the independent variable.

    Linear: Regression From data itself overall,consistent

    Logistic: Classification From the distribution point of view likelihood-based
23. (。。。超出能力范围了，我凭感觉乱写了)

    log of odds:

$ln\frac{p}{1-p}$

$ln\frac{h_\theta(x)}{1-h_\theta(x)} = ln\frac{p(y=1|x)}{p(y=0|x)} = \theta^T x$

which is a sigmoid function. the number cannot mean much but the threshold can be universal and much more normalized information.

And the entropy has the form:

$E[-log\ p] = -\sum p_i log\ p_i = -[plog\ p + (1-p)log\ (1-p)] = -plog\ \frac{p}{1-p} - log(1-p)$

- From a formal point of view, the two are also very related, so the log of odd is something related to the entropy and effective information

24. Conceptually: Convert the problem to a binary problem. Even can further divide the divided part .

    Computationally: Limit the result to 0-1. The computation is less complex. Efficient because does not need to calculate the exact number of x. More robust to the noises.

25. a. Generative methods try to model joint probabilistic distribution using Bayes formula, while discriminate methods try to model conditional probabilistic distribution. b. In general, generative methods requires much more training instances than discriminate methods, and thus suffers higher computational complexity. c. However, generative methods usually provide us with more insight into how data is generated. d. Discriminative methods can either have probabilistic interpretation or not. Generative models are purely based on probability theories.

    26,27是玄学问题。。。烧香拜佛，考了就自己发挥吧

26. Joint Distribution
27. Knowledge Graph
28. Key advantages of linear model: Such framework minimizes interactions between different factors, and also has very low computational complexity. Integration of linear models can solve complex problems.

    Not expressive: Linear regression is poor when the variables are nonlinear.And it is not flexible enough to capture more complex patterns, making it difficult and time consuming to add correct interaction terms or use polynomials.(y老师好像没怎么讲这个问题？反正就是一条线的表示能力可不有限么。。。。)

29. Key problems with complex NN: (1)It sustains the curse of combinatorial explosion such as network topology and a dramatic huge number of parameters. (2)Multiplication decreases the degree of convexity and therefore the model becomes more sensitive to the initial value.

    (感觉28,29好像重点集中在interaction上，好处就是interaction 少，坏处就是interaction多)

30. Solving its dual problem (Lagrange Multiplier)

    Find its equivalent problems (modify objective function)

    Using kernel tricks

31. Given an optimization problem：

    $\min_x f(x)$ s.t.$g_i(x) \leq 0, i = 1, \dots, k\ h_i(x) = 0, i = 1, \dots, l$

    We can get the Lagrange:

    $L(x, \alpha, \beta) = f(x) + \sum_{i=1}^k \alpha_i g_i(x) + \sum_{i=1}^l \beta_i h_i(x)$

    Prime problem: $\min\limits_x \max_{\alpha \geq 0, \beta} L(x, \alpha, \beta)$

    Dual problem: $\max_{\alpha \geq 0, \beta} \min_x L(x, \alpha, \beta)$

Strong Duality:For a minimization problem, denote the optimal value of the primal problem by p^*, and correspondingly the optimal value of the dual problem by d^*. We always have $d^* \leq p^*$. The strong duality means $d^* = p^*$.

32. (还是可以看拳师的note。。拳师用心良苦啊)

    Given an optimization problem：

    $\min_x f(x)$ s.t.$g_i(x) \leq 0, i = 1, \dots, k$ $h_i(x) = 0, i = 1, \dots, l$

We can get the Lagrange:

$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i g_i(x) + \sum_{i=1}^{l} \beta_i h_i(x)$

KKT conditions: for $\forall i \in \{1, \dots, k\}$

$\frac{\partial L}{\partial x} = 0, \frac{\partial L}{\partial \beta} = 0$ $g_i(x) \leq 0$ $\alpha_i \geq 0$ $\alpha_i g_i(x) = 0$

The original optimization problem is equivalent to optimizing its Lagrange function constrained by the KKT conditions. In the case of SVM, the optimization problem is:

$\min_{w,b} \|w\|$ s.t. $1 - w^T x_i + b \leq 0$

the KKT conditions are:

$$\begin{aligned} \alpha_i &\geq 0 \\ y_i \left(w^T x_i + b\right) - 1 &\geq 0 \\ \alpha_i \left(y_i \left(w^T x_i + b\right) - 1\right) &= 0 \end{aligned}$$

By solving its dual problem, we get $w = \sum_{i=1}^{m} \alpha_i y_i x_i$, then the final model:

$f(x) = w^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$

For any data point (x_i, y_i), if $\alpha_i = 0$, it won't appear in the final trained model. If $\alpha_i > 0$, then it's a support vector lying on the border of the maximum margin. Finally, only support vectors will appear in the formula for prediction, which implies the nature of sparsity of SVM.

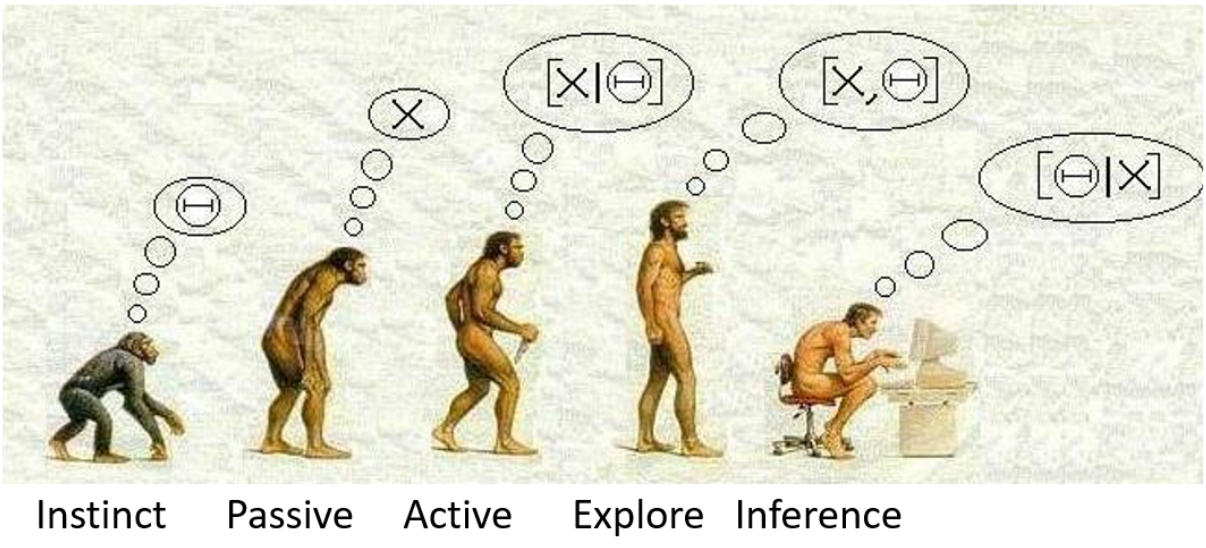
33. Soft margin SVM allows misclassication by introducing penalty on those misclassied cases. It improves the ability to tolerate noisy data and issues a model even when the problem is nonlinear. In soft margin SVM, the optimization object becomes:

$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l\left(y_i\left(w^T x_i + b\right) - 1\right)$

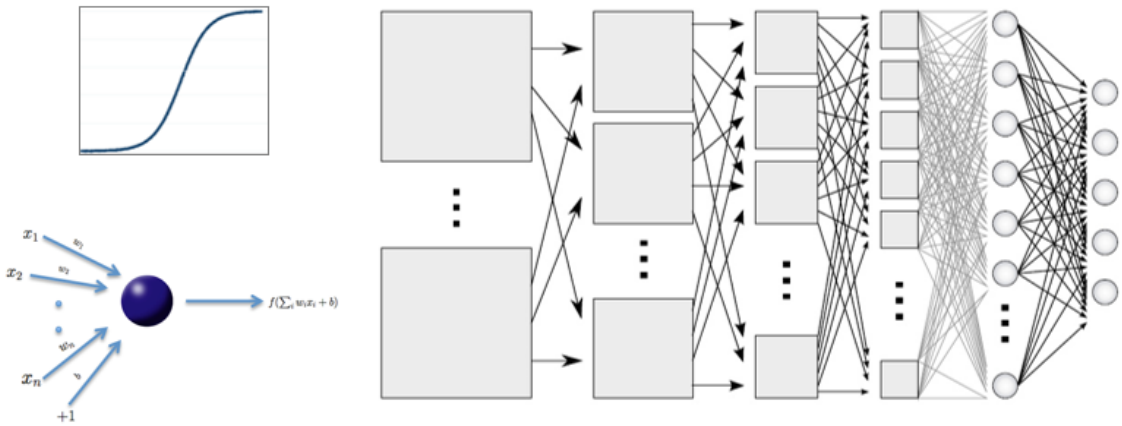
where $l(\cdot)$ is loss function and C is the penalty constant. $\frac{1}{2}\|w\|^2$ in the above formula can be regarded as a l_2 regularization term.

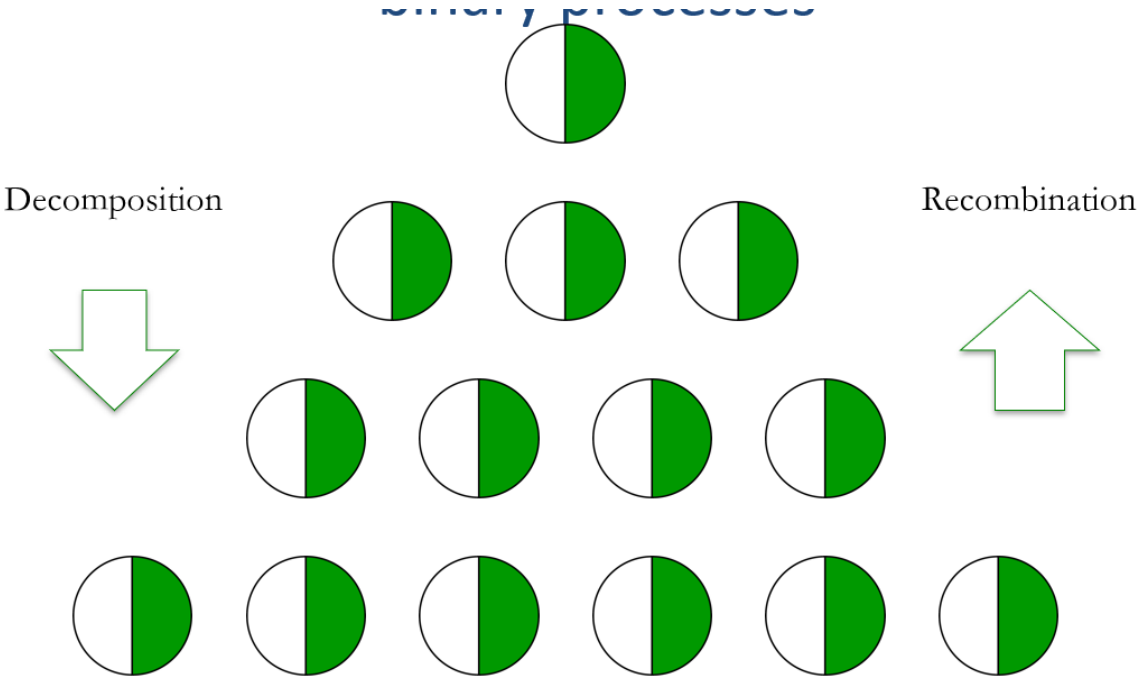
34. (1)Notice that in the entire computational process of SVM, x only emerges in the form of inner product $x_i^T x_j$. Therefore, we can introduce a kernel function $k(x_i;,x_j)$ to replace the original $< x_i, x_j >$. This is usually called kernel trick. (2)In essence, a kernel function corresponds to a mapping feature space. As we only replace $x_i^T x_j$ by $k(x_i;,x_j)$, rather than first mapping original data point to the new feature space and then computing the inner product, so there is not much additional computational complexity. (3) Since $k(x_i;,x_j))$ = $\phi(x)^T\phi(y)$, the kernel trick actually eliminates some common terms to reduce the computation. For example, the Gaussian Kernel $\exp\left\{-\frac{|x-y|^2}{2}\right\}$ actually decreases the computation from infinite many times to finite times.
35. • General idea: mapping data in the original feature space to a new feature space • Key computation: replace $x_i^T x_j$ by $k(x_i;,x_j)$ • locality, linearity and convexity. (etc. nonlinear to linear)
36. In many situations, the real data space is usually non-Euclidean. Therefore, we want to project a distance measured in Euclidean space to its real space. (nonlinear to linear)

    （玄学问题和个人理解问题建议自己查查看看听听吧，实在是太难整理了）
37. Recursion: logic and state of each recursion, how the states communicate with each other.(Things have to be propagated). Dependence of each recursion(within itself and between each other). Sequential? (他一直在说什么sequenlization)Parallel, Concurrence.
38. 大体意思就是不存在真正意义上的并行。用图灵机不可能实现并行。
39. 10th:Is there a general algorithm which, for any given Diophantine equation, can decide whether the equation has a solution for all unknowns with integer values.(whether a algorithm can lead to a solution,the limit of Turin machine,math,philosophy,answer is no)
40. 13th:Can a given high-dimensional problem be described by a composition with a finite number of bivariate functions?(motivation for anything. example: RNN support function)
41. we don't have answer for this! very open!
42. Recursion to update the prior and structure, search problem.(不知道为什么又开始bb41了，我感觉就是下面这张图吧)



43. 一切的基础都是prior
44. Hilbert10,13.The structural and logical recombination of binary processes





45. (homogeneous, orthogonal, normalized)The situation is too ideal.

    Multiple differential(depends on each other) three body problem.
46. (基本没讲为什么，就是函数复杂了导数就难求，初值敏感问题(initial sensitive))
47. Need a order which cannot be set. (还是没有真正意义上的并行)
48. 先说了反义词 heterogeneous-homogeneous; multiple scale-local/own

    mathematical heterogeneous: we need to integrate independent problems.

    Multiple scale: Combine all the local scales and normalize
49. (1)Composite is generally not convex but sometimes can be convex.

    Conditions:fixed/linear?

    $AB \cdot X$ fix one and only change one.

(2)We can use Relu as the activation function.

Capsule: divide it into partitions/regions which are convex.

50. Because there are many alternative way to relate X and Y.