1. **ANN: cost function, C(s,y):** När man fått ett värde på s (predicted output) så jämförs det med ett förväntat output (y). Detta sker m.h.a en cost function C(s,y). Cost can be equal to MSE, cross-entropy or any other cost function.
   Based on C's value, the model "knows" how much to adjust its parameters in order to get closer to the expected output y.
   This happens using the back-propagation algorithm.
2. **ANN: forward propagation:**

$x = a^{(1)}$ *Input layer*

$z^{(2)} = W^{(1)}x + b^{(1)}$ *neuron value at Hidde*

$a^{(2)} = f(z^{(2)})$ *activation value at Hidden*$_1$

$z^{(3)} = W^{(2)}a^{(2)} + b^{(2)}$ *neuron value at Hid*

$a^{(3)} = f(z^{(3)})$ *activation value at Hidden*$_2$

$s = W^{(3)}a^{(3)}$ *Output layer*

The "normal" ANN. The equations in the picture form network's forward propagation.

3. **ANN: what is backpropagation:** backpropagation aims to minimize the cost function by adjusting network's weights and biases. The level of adjustment is determined by the gradients (derivera och minimera vektorn C) of the cost function with respect to those parameters.
   Initialization of the weights in the backpropagation algorithm is crucial.
4. **ANN; why not use linear activation functions:** in that case the output would be a linear version of the input, meaning that all layers could be reduced to just one. Use sigmoid reLU or tanh.
5. **Approach to model selection:**
   - Given several models $M_1, \dots M_m$
   - Divide data set into **training** and **test** data

Training	Test

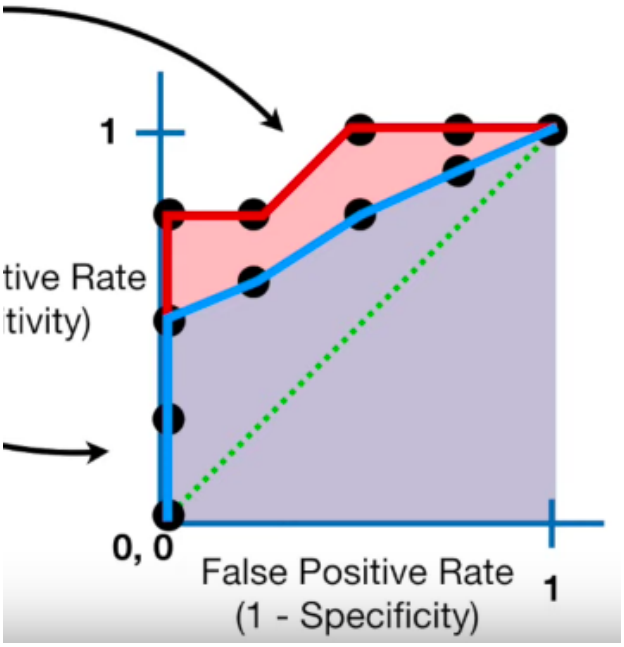
   - Fit models M_i to training data→get parameter values
   - Use fitted models to predict test data and compare **test errors** $R(M_1), \dots R(M_m)$
   - Model with lowest prediction error is best
6. **The assumption in naive bayes classifier:**

$$P(X = (x_1, \dots x_p)|Y = y) = \prod_{i=1}^{p} P(X_i = x_i|Y = y)$$

The assumption is that the predictors/features are independent That is presence of one particular feature does not affect the other. Hence it is called naive.

Another assumption made here is that all the predictors have an equal effect on the outcome.

7. **AUC:**



AUC stands for area under the curve, and is the area under the ROC-graph. It can be used to compare classification methods, higher AUC is better.

8. **Basic ML ingredients:**

- **Data** D: observations (cases)
  - Features $X_1, .. X_p$
  - Targets $Y_1, \dots, Y_r$

Case	X_1	
1		
2		
...		

- **Model** $P(x|\, w_1, \dots w_k)$ or $P(y|x, w_1, \dots w_k)$
  - Example: Linear regression $p(y|x, w) = N(w_0 + w_1 x, \sigma^2$
- **Learning procedure** (data→get parameters $\hat{w}$ or p
  - Maximum likelihood, Bayesian estimation...
- **Prediction** of new data X^{new} by using the fitted m

Data, Model, Learning procedure, Prediction.

9. **Bayes theorem:**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

a method used to compute posterior probabilities

10. **Big advantage of Support Vector Machines (SVM) and Logistic Regression over K nearest neighbors:** Support Vector Machines (SVM) and Logistic Regression trys to find a line to separate the data, once the line/plane/hyperplane is found, new data points can be very quickly classified.

11. **The big difference between ridge and lasso regression:** Ridge regression can only shrink the slope asymptotically close to zero while lasso can shrink the slope all the way to zero. Så Lasso är lite bättre när vissa variabler är överflödiga, och Ridge är lite bättre när de flesta variabler är användbara.

12. **CART algorithm for finding optimal regression tree. (fitting):**

1. Let C_0 be a hypercube containing all observations
2. Let queue C={C_0}
3. Pick up some C_j from C and find a variable X_j and value s tha two hypercubes

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j$$

and solve

$$\min_{j,s}[N_1 Q(R_1) + N_2 Q(R_2)]$$

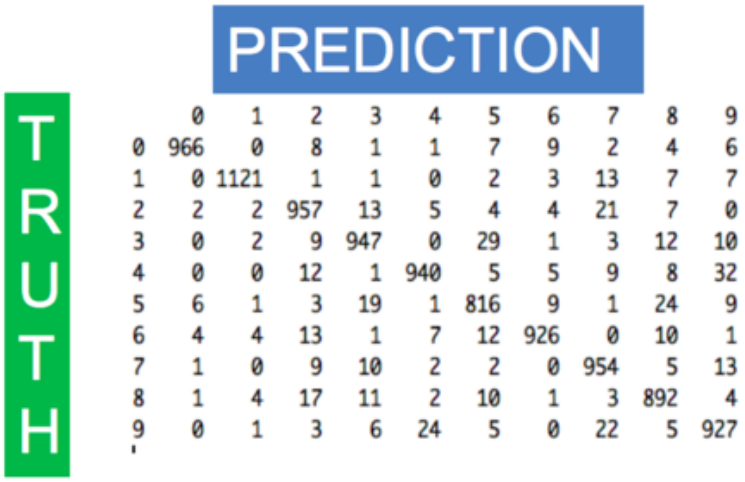
4. Remove C_j from C and add R_1 and R_2
5. Repeat 3-4 as many times as needed (or until each cube has observation)

Its a greedy algorithm (optimal solution not found)

13. **Choosing variables: Forward stepwise selection:**

- Starts with 0 features (or full set) and then adds a feature (removes feature) that most improves the measure selected.
  - Can handle large p quickly
  - Does not examine all possible subsets (not the "best")

14. **classifying hadwritten digits, example Confusion Matrix:**



	0	1	2	3	4	5	6	7	8	9
0	966	0	8	1	1	7	9	2	4	6
1	0	1121	1	1	0	2	3	13	7	7
2	2	2	957	13	5	4	4	21	7	0
3	0	2	9	947	0	29	1	3	12	10
4	0	0	12	1	940	5	5	9	8	32
5	6	1	3	19	1	816	9	1	24	9
6	4	4	13	1	7	12	926	0	10	1
7	1	0	9	10	2	2	0	954	5	13
8	1	4	17	11	2	10	1	3	892	4
9	0	1	3	6	24	5	0	22	5	927

How many picture 2 are predicted as for example a 5?

15. **Confindence intervals in regression:**

Estimation

1. Compute $D_1, \dots D_B$ using a bootstrap
2. Fit model to $D_1, \dots D_B$ →estimate $\hat{W}_1, \dots \hat{W}_B$
3. For a given X, compute $f(X, \hat{w}_1), \dots f(X, \hat{w}_B)$ and estimate confidence interval by (percentile method)
4. Combine confidence intervals in a band

16. **conjugate distributions:** if the posterior distributions p(θ I x) are in the same probability distribution family as the prior probability distribution p(θ), the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Conjugate distributions -> posterior is also that distribution.

17. **cross validation:** en test metod som går ut på att använda olika delar av datat som test och sedan sätta ihop resultatet

18. **degrees of freedom:** degrees of freedom (DF) indicate the number of independent values that can vary in an analysis without breaking any constraints.

19. **Deterministic vs probabilistic model:** A deterministic mathematical model is meant to yield a single solution describing the outcome of some "experiment" given appropriate inputs. A probabilistic model is, instead, meant to give a distribution of possible outcomes (i.e. it describes all outcomes and gives some measure of how likely each is to occur).

20. **Downsides with decision trees:**

- Trees have high variance: a small change in respon different tree
- Greedy algorithms → fit may be not so good
- Lack of smoothness

21. **Entropy / Deviation as impurity measure:**

$$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

22. **Example K-nearest neighbor density estimation:**

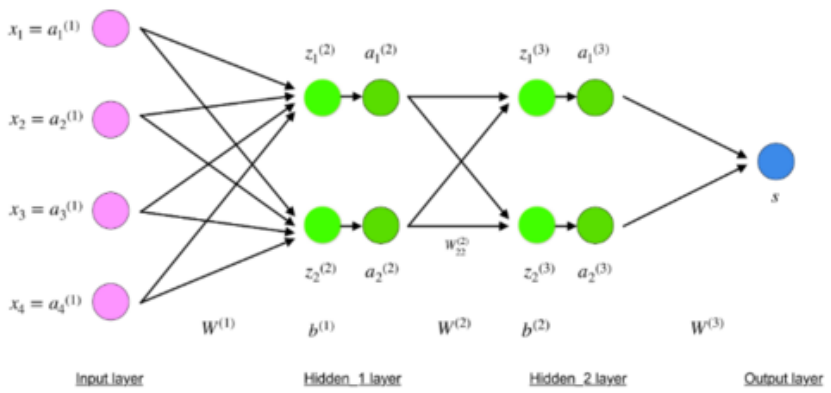
- Data: Fish length $X_1, \ldots X_N$
- Model $p(x|K) = \frac{K}{N \cdot \Delta}$
  - K: #neighbors in training data
  - Δ: length of the interval containing K nei
- Learning: Fix some K or find an appro
- Prediction: predict $p(x|K)$

N = total data. Delta = volym av område(?). K = alla data inne i område.

23. **Example Logistic regression:**

- Data $Y_i \in \{Spam, Not\ Spam\}, X_i = \#of$
- Model: $p(Y = Spam|w, x) = \frac{1}{1+e^{-w_0-w_1}}$
- Fitting: maximum likelihood
- Prediction : $p(spam) = p(Y = spam|x)$

24. **exemple ANN:**



Simple 4-layer neural network illustration

a = activations, uträknade av en activation function.

z2 = weight matrix * x + bias.

a2 = f(z2)

z3 = weight matrix * a2 + bias

25. **Exponential family of distributions:**

- Normal, Exponential, Gamma, Beta, Chi-squared..
- Bernoulli, Multinoulli, Poisson...

Non exponential -> uniform, Student t

26. **Finding best classification from a classification tree.:**

$$k(m) = \arg\max_k \hat{p}_{mk}$$

Classification probability pmk=p(Y=k | X ∈ Rm) is estimated for every class in a node .

27. **From kernel density estimation to kernel classification:**

1. Estimate $p(\boldsymbol{x}|y = 0)$ and $p(\boldsymbol{x}|y = 1)$ using the methods just seen.
2. Estimate $p(y)$ as class proportions.
3. Compute $p(y|\boldsymbol{x}) \propto p(\boldsymbol{x}|y)p(y)$ by Bayes theorem.

28. **Functions for calculating error, regression vs classification:**

- Regression, **MSE** :

$$R(Y,\hat{Y}) = \frac{1}{N}\sum_{i=1}^{N}(Y_i -$$

- Classification, **misclassification**

$$R(Y,\hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i =$$

29. **Generalized Linear Model:**

Model	Random	Link
Linear Regression	Normal	Identity
ANOVA	Normal	Identity
ANCOVA	Normal	Identity
Logistic Regression	Binomial	Logit
Loglinear	Poisson	Log
Poisson Regression	Poisson	Log
Multinomial response	Multinomial	Generalized Logit

The term generalized linear model (GLIM or GLM) refers to a larger class of models. It is like linear regression but also counts with distribution of dependent variable and a link function. Link function makes up for that is that the effect of the predictors on the dependent variable may not be linear in nature.

The .glm() function is the basic tool for fitting generalized linear models.

30. **The Gini Index:**

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Mått på impurity av en nod. The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. (så tex 1- (36/70)^2 - (34/70)^2, för att 36 observationer i noden var klass 'yes' och 34 'no' angående heart disease. It favors larger partitions. När noden är pure så är gini index 0.
Man vill ha ett lågt Gini index.

31. **Histogram Classification:**

- divide the input space into D-dimensional cubes of side h, and
- classify according to majority vote in the cube $C(\boldsymbol{x}, h)$ that contains $\boldsymbol{x}$.

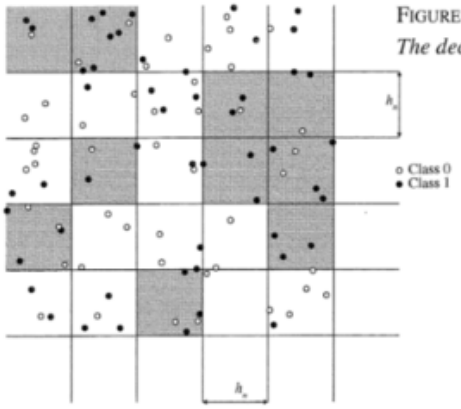
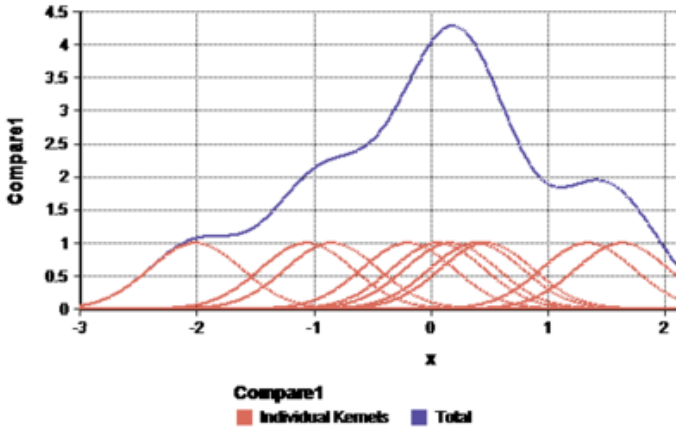


FIGURE 6.1. *A cubic histogram rule: The decision is 1 in the shaded area.*

D = number of inputs?
The histogram rule is less accurate at the borders of the cube, because those points are not as well represented by the cube as the ones near the center. (moving window classification is better at this)

32. **How estimate uncertainty of estimator f? (estimator of probabilistic model f(x,D):** 1. If data D has a known distribution, compute distribution for estimator (difficult)
2. Use Bootstrap method.
33. **How find good penalty factor for Ridge?:** Testa modell med olika värde på lambda (penalty factor) på training data och välj den med minst error. (cross-validation)
34. **How makes LDA a prediction:** LDA makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made. The model uses Bayes Theorem to estimate the probabilities.
35. **How to model bernoulli or multinomial targets ?:** Logistic regression
36. **How to model normally distributed targets:** Linear regression
37. **Hur välja rot node:** Välj den som har lägst totalt impurity. I alla lägen i trädet placera noden med lägst impurity.
38. **hyperparameter:** A hyperparameter is a parameter whose value is set before the learning process begins. By contrast, the values of other parameters are derived via training. Tex; lambda i Ridge och Lasso. Eller width i Kernel.
39. **ICA (independent component analysis):** Making a linear transformation such that the features becomes independent of eachother while maximizing the similarity to the original features.

40. **Kernel Density Smoothing:**



Each sample point is replaced by a Gaussian Kernel (red curves), then they are

Replaces each sample point with a Gaussian-shaped Kernel, then obtains the resulting estimate for the density by adding up these Gaussians. The width (h) of the kernels can be determined with cross-validation. Smoothing increases variance.

41. **Kernel function:** The kernel function is what is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable. (kernel trick)

42. **Lasso:**

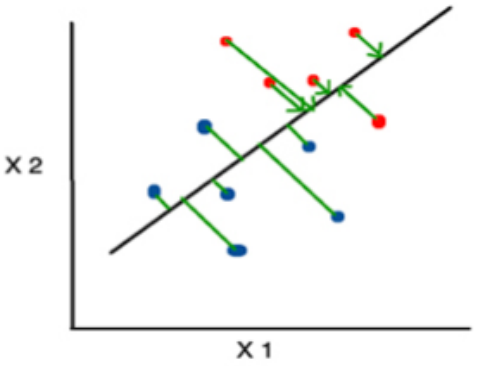
$$\hat{w}^{lasso} = \text{argmin}\left\{\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1j} - \ldots - w_p x_{pj})^2 +\right.$$

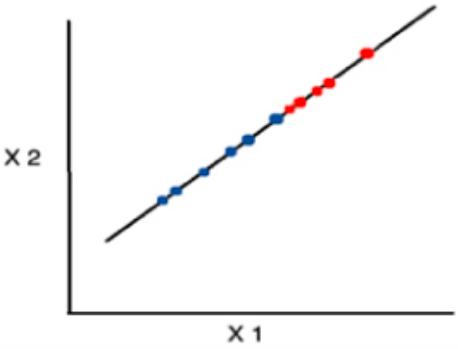
Like Ridge but with L1 regularization instead of L2.

43. **LDA makes some simplifying assumptions about your data.:**
That each variable is Gaussian.
That each attribute has the same variance.

With these assumptions, the LDA model estimates the mean and variance from your data for each class.

44. **LDA projecting the features in higher dimension space to lower dimensions, visualized:**





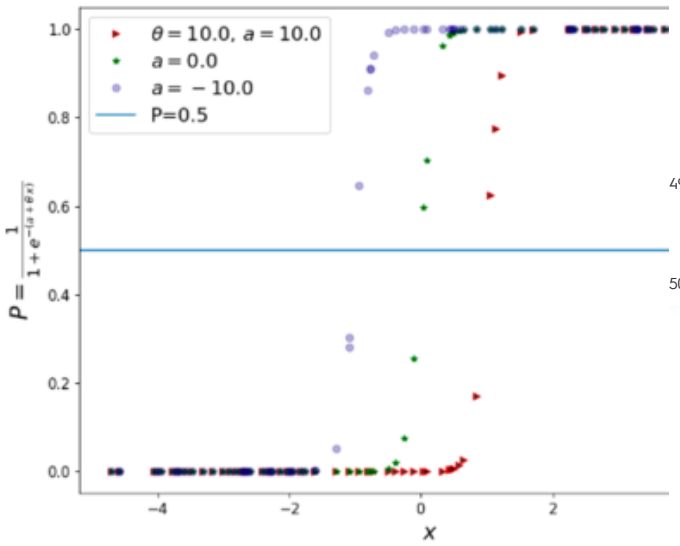
From this line the LDA model is able to estimate the mean and variance from your data for each class with the help of the assumptions that the data is gaussian with same variance.

Outliers can skew the primitive statistics used to separate classes in LDA, so it is preferable to remove them.

Since LDA assumes that each input variable has the same variance, it is always better to standardize your data before using an LDA model. Keep the mean to be 0 and the standard deviation to be 1.

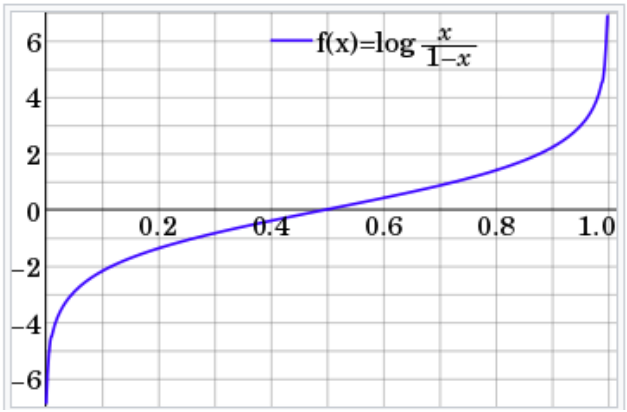
45. **logistic regression:** Logistic regression is useful when predicting binary outcomes from continuous predictor variables. (Example, if a loan is accepted or denied)

46. **logistic regression inner functioning:**



Output är binärt. Linear regression är output [-infinity,infinity]. Man måste göra om sannolikheterna från [0,1] till [-infinity,infinity], detta görs med en link funktion. Om man sedan löser ut p så får man följande graf. Man kan nu göra klassifiering. (ger bättre resultat än att inte använda family / link function, kanske för att man optimerar parametrar theta och a)?

47. **logit, a link function for a generalized linear model:**



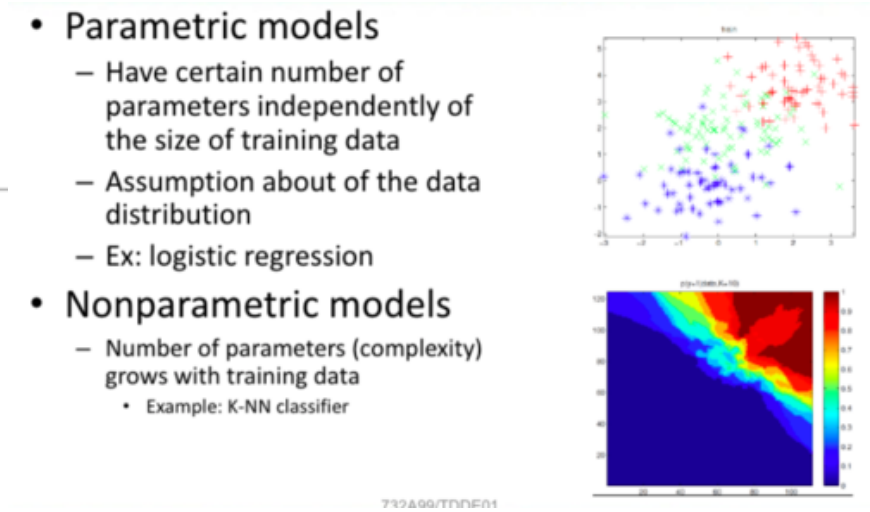
(pictures shows plot of logit where x is probability, the link function for binomial family. It maps probabilities [0,1] to [-infinity,infinity]). Logit is the canonical link function for the bernoulli distribution.

(x/1-x istället för y => [0,1] blir till [0, infinity], log på detta för att få [-infinity,infinity])

48. **maximal margin classifier:** Finds a threshold between in the middle between observations to divide the data. This classifier has a hard margin and is very sensitive to outliers. It wont allow missclassifications in the sense that if one observation gets missclassified it can move the threshold to a non optimal position. The threshold gets updated when new observations are made.

49. **mean squared error (MSE):** measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

50. **Model Types:**



51. **Moving Window Classification:**

- consider the points within a certain distance to the point to classify, and
- classify the point according to majority vote.

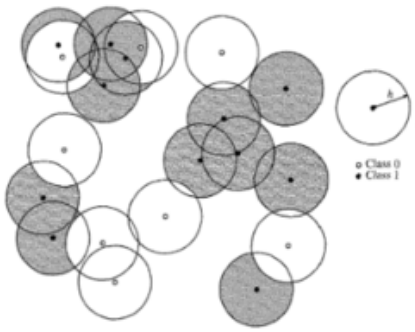
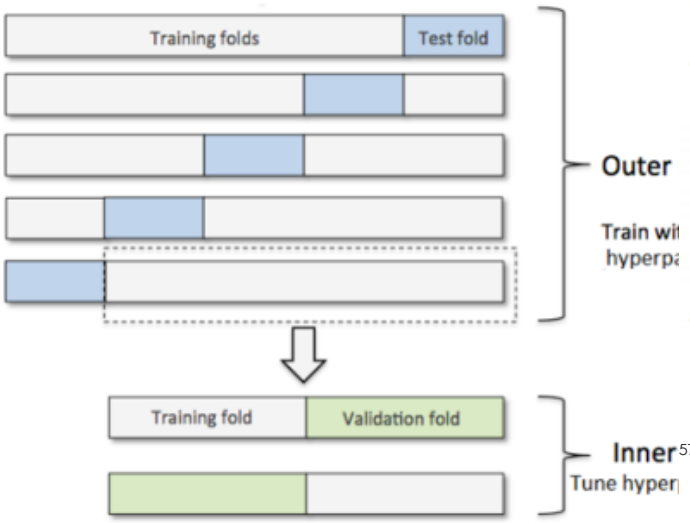


FIGURE 10.1. *The moving window rule in $\mathcal{R}^2$. The decision is 1 in the shaded area.*

The moving window rule gives equal weight to all the points in the ball.

52. **MSE criterion:** MSE criterion is a tradeoff between bias and variance. It is important especially when the polynomial has higher degree to avoid overfitting.

53. **Nested cross-validation.:**



54. **Nonparametric bootstrap:**

Given estimator $\hat{w} = \hat{f}(D)$

Assume $X \sim F(X, w)$, F and w are unknown

1. Estimate $\hat{w}$ from data $\mathbf{D}=(X_1,\ldots X_n)$
2. Generate $\mathbf{D_1} = (X^*_1,\ldots X^*_n)$ by sampling with rep
3. Repeat step 2 B times
4. The distribution of w is given by $\hat{f}(D_1), \ldots \hat{f}(D$

can be applied to any deterministic estimator

55. **Ordinary least square regression, vad är RSS?:**

Estimation: maximizing the likelihood

$$\hat{w} = \max_{w} p(D|w)$$

Is equivalent to minimizing

$$RSS(w) = \sum_{i=1}^{n} (Y_i - \boldsymbol{w}^T \boldsymbol{X_i})^2$$

The sum of all the squared residuals (deviations) is known as the residual sum of squares (RSS) and provides a measure of model-fit for an OLS regression model.

56. **Parametric bootstrap:**

Given estimator $\hat{w} = \hat{f}(D)$

Assume $X \sim F(X, w)$, F is known and w is unknown

1. Estimate $\hat{w}$ from data $\mathbf{D}=(X_1,\ldots X_n)$
2. Generate $\mathbf{D_1} = (X^*_1,\ldots X^*_n)$ by generating from $F(X, \hat{w})$
3. Repeat step 2 B times
4. The distribution of w is given by $\hat{f}(D_1), \ldots \hat{f}(D_B)$

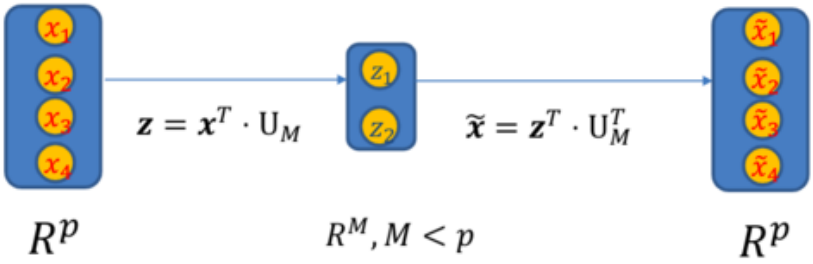
Works even for small samples.

57. **PCA i andra ord:** PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, such as that the variance is maximized.

Skapar oberoende principal components utav (möjligtvis) beroende features. Målet är att dessa principal components ska beskriva så mycket som möjligt av datan. En principal component är inte en feature. PCA removes correlations, but not higher order dependence. ICA removes correlations and higher order dependence.

In layman terms PCA helps to compress data and ICA helps to separate data

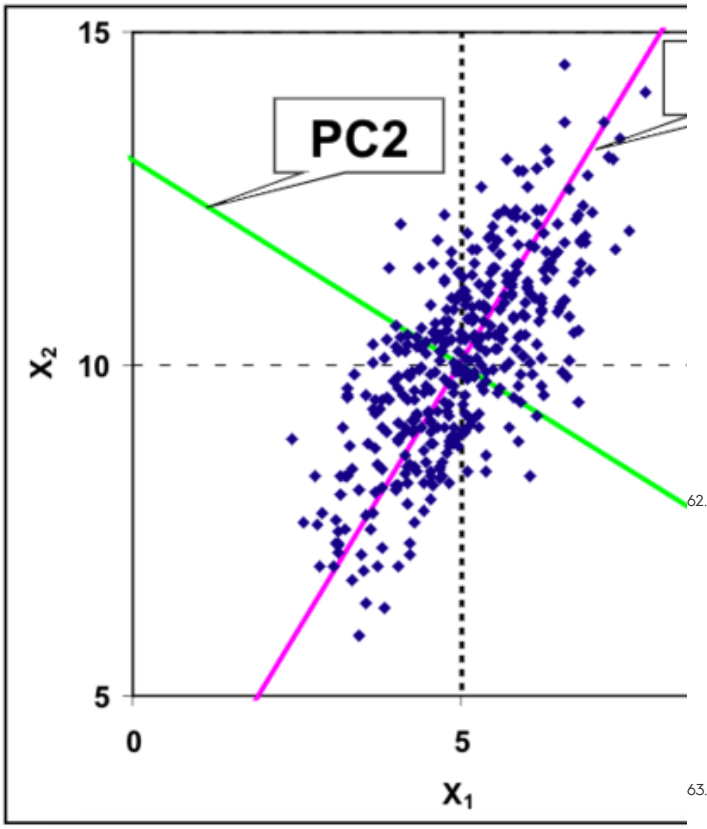
58. **PCA (principal component analysis):**



Reduces complexity in data, approximates data with fewer dimensions.

Finding correlation by maximizing variance.

59. **PCA process:**



Välj två eller fler features som axlar. Sätt ut datat.
Skapa en linje som minimerar avståndet mellan data och linje. Detta är PC1. Lutningen säger viktigheten av en feature. (tex k = 4 innebär att F1 är mkt viktigare än F2).
Lägg till en linje ortogonal mot PC1, detta är PC2.
Enhetsvektorn i PC1s riktning är eigenvektorn för PC1.
Eigenvärdet är summan av alla kvadrater.

60. **Praktisk samband mellan MLE (maximum likelihood estimation) och normalfördelning:** It turns out that when the model is assumed to be Gaussian, the MLE estimates are equivalent to the ordinary least squares method.

61. **probability density function:**

$$p(x|\theta) = \theta e^{-\theta x}$$

Any given sample can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample. In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there are an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample. (eller att i just den punkten är det viss sannolikhet / enhet)

62. **Projection matrix P (hat matrix):**

$$\hat{y} = X\hat{w} = X(X^T X)^{-1} X^T y = Py$$

Maps the vector of response values (dependent variable values) to the vector of fitted values (or predicted values). It describes the influence each response value has on each fitted value.

63. **Regularization:**

$$\hat{w}^{ridge} = \text{argmin}\left\{\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1j} - \ldots - w_p x_{pj})^2 + \lambda\sum_{j=1}^{p} w_j^2\right\}$$

Regularization works by biasing data towards particular values (such as small values near zero). The bias is achieved by adding a tuning parameter to encourage those values.

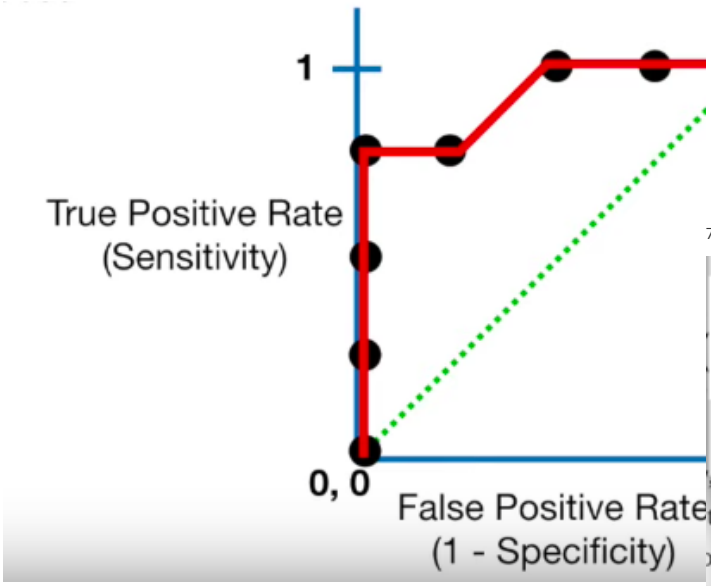
L2 regularization (ridge regression) adds an L2 penalty equal to the square of the magnitude of coefficients.

64. **Ridge regression idea:**

$$\hat{w}^{ridge} = \text{argmin}\left\{\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1j} - \ldots - w_p x_{pj})^2 + \lambda\sum_{j=1}^{p} w_j^2\right\}$$

Keeps all predictors of mean square but shrink coefficients to make model less complex. (med hjälp av L2 regularization)
Shrinking enables estimation of regression coefficients even if the number of parameters exceeds the number of cases.

65. **ROC graph:**



The ROC graph summarizes all of the confusion matrix that each threshold produces (in classification problem). Each dot represents a threshold. Y-axis is the ratio of positives classified as positive. X-axis is the ratio of negatives classified as positives.

66. **sampling with replacement:** Samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. (Used in Bootstrap)

67. **Second order dependency vs higher order dependency.:** Second order dependency is dependency between two variables, Cov(x,y) != 0. Higher order means dependency between >2 variables. Cov(x,y,z) != 0.

68. **sparse solution:** the majority of x's components (weights) are zeros, only few are non-zeros. And a sparse solution could avoid over-fitting.

69. **stepAIC and model selection:** If we are given two models then we will prefer the model with lower AIC value. Hence we can say that AIC provides a means for model selection. AIC is only a relative measures among multiple models.

70. **support vector classifier:** like maximal margin classifier but with a soft margin. (also called soft margin classifier). Bias/variance tradeoff, it allows some missclassifications to lower the variance. Observations inside the soft margin is called support vectors. The soft margin is determined with cross validation.

71. **support vector machine:** support vector classifiers cant handle data that is not easily divided by a single line/plane, but support machine vectors can.

    1. Begin with data in a lower dimension.
    2. Transpose the data to a higher dimension.
    3. Find a support vector classifier to separate the data in the new dimension in two groups. (with help of kernel functions)

72. **Tekniken bakom naive bayes:**

$$P(y|x_1,...,x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

1. Använd bayes theorem för att få fram ett uttryck för sannolikheten att en viss klass utifrån observerade värden.
2. Med bayes assumption får vi produkt av termer i nämnaren.
3. Eftersom att nämnaren är konstant så gäller att VL är proportionell med täljaren i HL.
4. Maximera HL på y för att få den mest sannolika klassen.

73. **TPR vs FPR:** TPR (true positive rate) är antal TP genom totalt antal positiva.

    FPR (false positive rate) är antal FP genom totalt antal negativa.

74. **Types of learning: Reinforcement learning.:** Find suitable actions to maximize the reward. True targets are discovered by trial and error

75. **Types of learning: Semi-supervised:** targets are known only for some observations.

76. **Types of learning: Supervised learning:**

- Compute parameters from data
- Given features of a new object, predict target
- **Classification** (Y=categorical), **Regression** (Y=continuous)

77. **Types of learning: Unsupervised learning:**

- Unsupervised learning (→Data Mining)
  - No target
  - Aim is to extract interesting informati
    - Relations of parameters to each other
    - Grouping of objects

78. **Upsides with decision trees:**

- Easy to handle all types of features in on
- **Automatic variable selection**
- Relatively robust to outliers
- Handle large datasets

79. **Vad är maximum likelihood estimation till för?:** MLE can be defined as a method for estimating population parameters (such as the mean and variance for Normal, rate (lambda) for Poisson, etc.) from sample data such that the probability (likelihood) of obtaining the observed data is maximized.

    we can assume that we have a likelihood function L(θ;x), where θ is the distribution parameter vector and x is the set of observations.

80. **Vad är sant om proportionalitet:**

$$P(y|x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} F$$

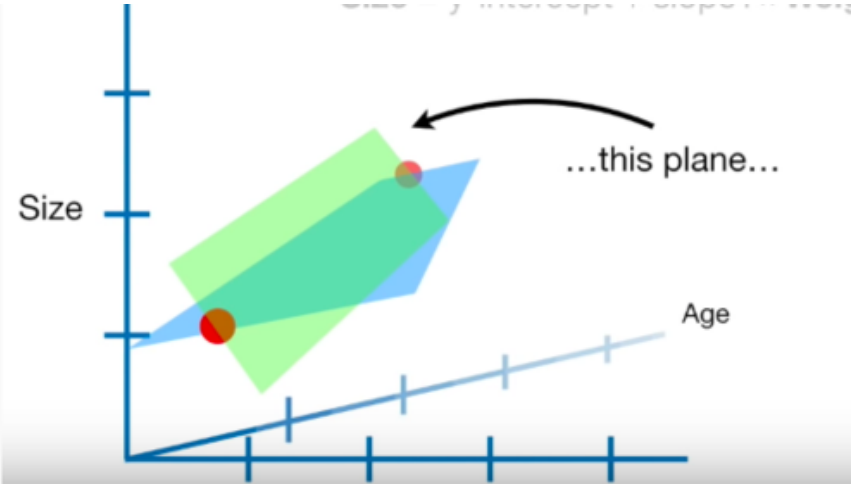
Att om VL och HL är proportionella så kommer det vara samma värde på Y som maximerar / minimerar uttrycken.

81. **Vad säger p(xIθ)*p(θ)?:**

$$p(Y|X) \propto p(X|Y)p(Y)$$

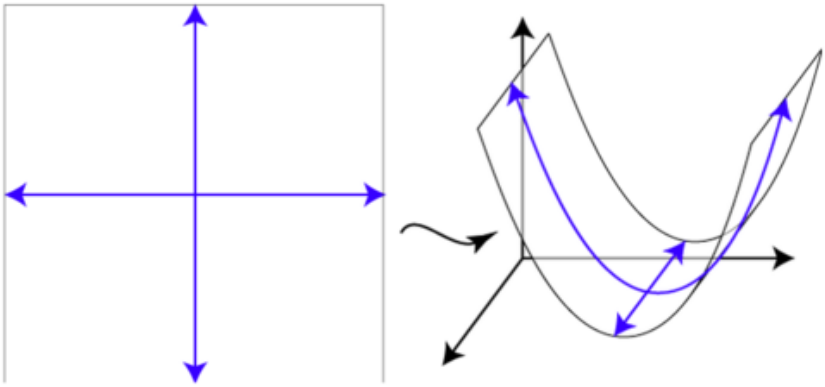
Eftersom att p(xIθ)*p(θ) är proportionell mot p(θIx), så kan

82. **varför med least squares metoden måste man ha minst lika många datapunkter som features:**



Exempel, tre features två datapunkter. Det spelar då ingen roll vilken lutning planet har i en av riktningarna, alltså går det inte att göra en prediction. Med ridge regression är detta dock möjligt

83. **What are Kernels used for:**



To generalize any linear algorithm to use curved shapes. (a function from the low dimensional space into a higher dimensional space.) The goal of the kernel is to make it so that two classes of data points that can only be separated by a curved line in the two-dimensional space can be separated by a flat plane in the three-dimensional space.

84. **What is stepAIC:** stepAIC is one of the most commonly used search method for feature selection. We try to keep on minimizing the stepAIC value to come up with the final set of features. "stepAIC" does not necessarily means to improve the model performance, however it is used to simplify the model without impacting much on the performance. So AIC quantifies the amount of information loss due to this simplification

85. **when consider the bayesian approach (use of bayesian model):** "I consider Bayesian approach when my data set is not everything that is known about the subject, and want to somehow incorporate that exogenous knowledge into my forecast"

    When there is uncertainty, and there might be relevant info to get from the prior, which would increase certainty.

86. **When is a node pure / impure:** Om dess alla observationer faller inom en klass. Tex alla som har chest pain -> headache -> high blood pressure, har heart disease. (faller inom klassen 'yes'). då är noden pure.

    Mäter impurity med gini index eller entropy/deviation

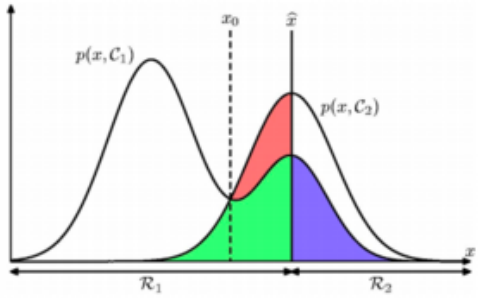
87. **When use LDA (linear discriminant analysis) and when use logistic regression for classification?:** Logistic regression is intended for two-class or binary classification problems. If you have more than two classes then Linear Discriminant Analysis is the preferred linear classification technique.

    It might though be a good idea to try both logistic regression and linear discriminant analysis.

88. **Why GLM instead of OLS regression?:** Ordinary Least Squares regression provides linear models of continuous variables. However, much data of interest to statisticians and researchers are not continuous and so other methods must be used to create useful predictive models. The glm() command is designed to perform generalized linear models (regressions) on binary outcome data, count data, probability data, proportion data and many other data types.

89. **Why making a density estimation might be interesting:**

1. Estimate **class-conditional densities** $p(x|y = C_i)$
2. Predict



90. **why ridge regression instead of least square?:** Least squares regression isn't defined at all when the number of predictors exceeds the number of observations; It doesn't differentiate "important" from "less-important" predictors in a model, so it includes all of them. This leads to overfitting a model and failure to find unique solutions. Ridge forces the model to use fewer predictors.