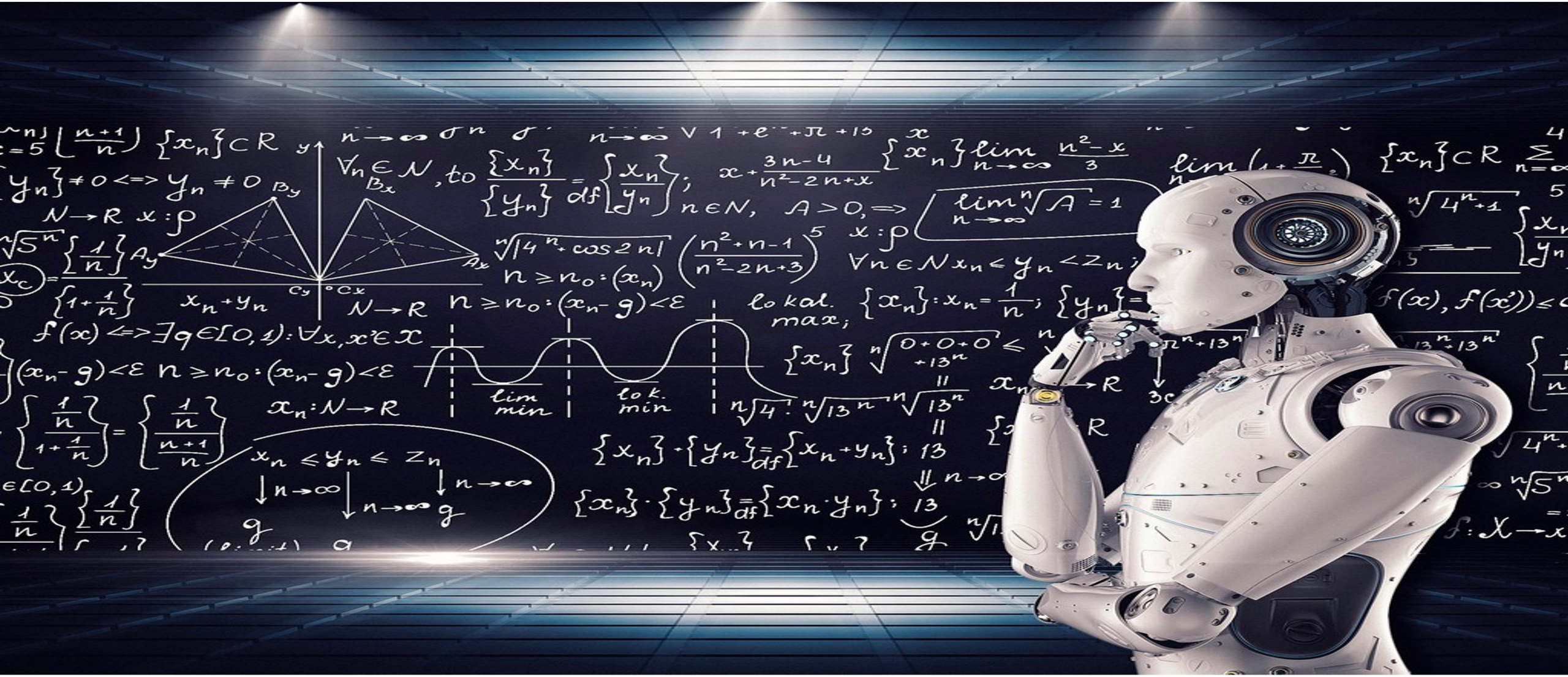
Bayesian Learning



Appeals procedure

- After the grades are published, you have one week to send a mail with your appeal
- You will get a response mail with the appeal's decision

Probability algorithms

- Our previous model didn’t use probability calculation (except maybe in the goodness of split)
- The most intuitive algorithm is to return the majority class, or in other word – return the most probable class according to the training set
- Today agenda – probability algorithms – algorithms that uses probability techniques in order to predict a new instance target value/class.

Probability recap

- Sample space
- A sample space is a **set** of events which lists all possible outcomes:
  - For a coin toss this is the sample space: S = {H,T}
  - For rolling a dice this is the sample space: S = {1,2,3,4,5,6}
  - For rolling two dice this is the sample space: S = {(1,1), (1,2), (1,3), (1,4), …, (6,5), (6, 6)}
  - For a person Height this is the sample space S = [0, 3]

Probability recap

- Events
- Any subset of the sample space is called an event
  - For rolling two die an example event could be:
    E={(1,6) , (2,5), (3,4), (4,3), (5,2), (6,1)}

- Some basic operation on events:
  - Union
  - Intersection
  - Complement

Probability recap

- Random variable
- Some function of the outcome event $X: \Omega \to \mathbb{R}$

- For example, the sum of two die (not the two numbers that come up):
  - Let *X* be a random variable denoting the sum of two dice rolls
    - $X(\{1,3\}) = 4$
    - X({6,5}) = 11

  - And we can ask what is the probability of seeing a specific X
    - $P(X = 1) =$?
    - $P(X = 2) = P(\{1,1\}) =$?
    - $P(X = 4) = P(\{1,3\},\{3,1\},\{2,2\}) =$?

Probability recap

- Random variable
- We now can define the expected value of a random variable:
  - For discrete variable:

$$E[X] = \sum_{x} xp(x)$$

  * Where p is the probability mass function (pmf)
  - For continuous variable:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

  * Where f is the probability density function (pdf)

Probability recap

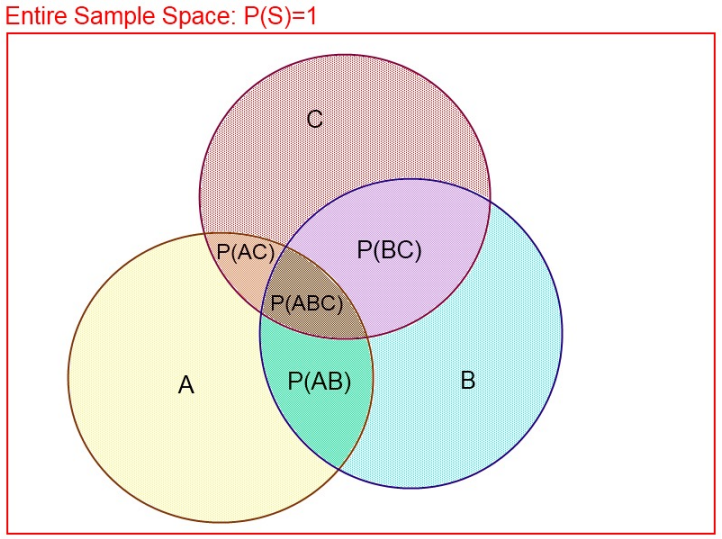
- Random variable
- The variance:

$$\sigma^2 = var(X) = E[(x-\mu)^2]$$

- The standard deviation (=square root of the variance):

$$\sigma = \sqrt{var(X)} = \sqrt{E[(x-\mu)^2]}$$

Probability recap



- $P(A \cup B) = ?$

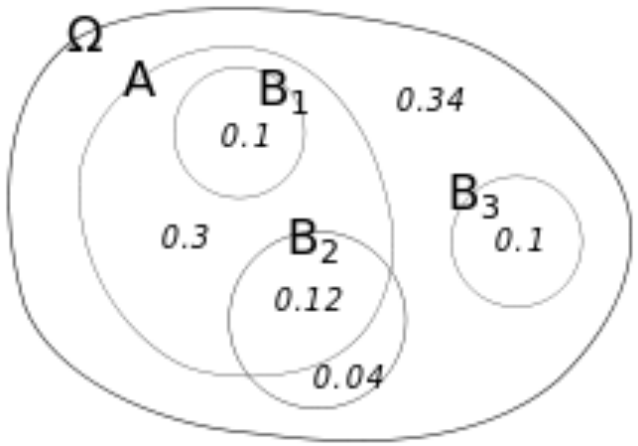
$$P(A) + P(B) - P(A \cap B)$$

- $P(A \cup B \cup C) = ?$

$$P\big((A \cup B) \cup C\big) =$$
$$P(A \cup B) + P(C) - P\big((A \cup B) \cap C\big) =$$
$$P(A) + P(B) - P(A \cap B) + P(C) - P\big((A \cap C) \cup (B \cap C)\big) =$$
$$P(A) + P(B) - P(A \cap B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) =$$
$$P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Probability recap

- Conditional probability:
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - $P(A|B_1) = ?$
    - $\frac{P(A \cap B_1)}{P(B_1)} = \frac{0.1}{0.1} = 1$
  - $P(A|B_2) = ?$
    - $\frac{P(A \cap B_2)}{P(B_2)} = \frac{0.12}{0.16} = 0.75$
  - $P(A|B_3) = ?$
    - $\frac{P(A \cap B_3)}{P(B_3)} = \frac{0}{0.1} = 0$



Probability recap .

- Example:
  - $P(Pass) = 90\%$
  - $P(Fail) = 10\%$

- We also know that:
  - $P(Learn\ for\ the\ test|Pass) = 90\%$
  - $P(Didn't\ learn|Pass) = 10\%$
  - $P(Learn\ for\ the\ test|Fail) = 5\%$
  - $P(Didn't\ learn|Fail) = 95\%$

- What is the probability that you pass the test if you learn?

Probability recap

- $P(Pass \cap Learn\ for\ the\ test) = P(Pass) \times P(Learn\ for\ the\ test|Pass)$
  $= 90\% \times 90\% = 81\%$
- $P(Pass \cap Didn't\ learn) = P(Pass) \times P(Didn't\ learn|Pass) = 90\% \times 10\%$
  $= 9\%$
- $P(Fail \cap Learn\ for\ the\ test) = P(Fail) \times P(Learn\ for\ the\ test|Fail)$
  $= 10\% \times 5\% = 0.5\%$
- $P(Fail \cap Didn't\ learn) = P(Fail) \times P(Didn't\ learn|Fail) = 10\% \times 95\%$
  $= 9.5\%$
- $P(Learn\ for\ the\ test) =$?

- $P(Didn't\ learn) =$?

Probability recap

- $P(Pass|Learn\ for\ the\ test) = \frac{P(Pass \cap Learn\ for\ the\ test)}{P(Learn\ for\ the\ test)}$
- $P(Fail|Learn\ for\ the\ test) = \frac{P(Fail \cap Learn\ for\ the\ test)}{P(Learn\ for\ the\ test)}$
- $P(Pass|Didn't\ learn) = \frac{P(Pass \cap Didn't\ learn)}{P(Didn't\ learn)}$
- $P(Fail|Didn't\ learn) = \frac{P(Fail \cap Didn't\ learn)}{P(Didn't\ learn)}$

Probability recap

- Independent events
  - If $P(A \cap B) = P(A)P(B)$ then A & B are independent
  - From conditional probability we get:

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

$$\Downarrow$$

$$P(A \cap B) = P(A|B)P(B)$$

  - If A & B are independent:

$$P(A)P(B) = P(A \cap B) = P(A|B)P(B)$$
$$P(A) = P(A|B)$$

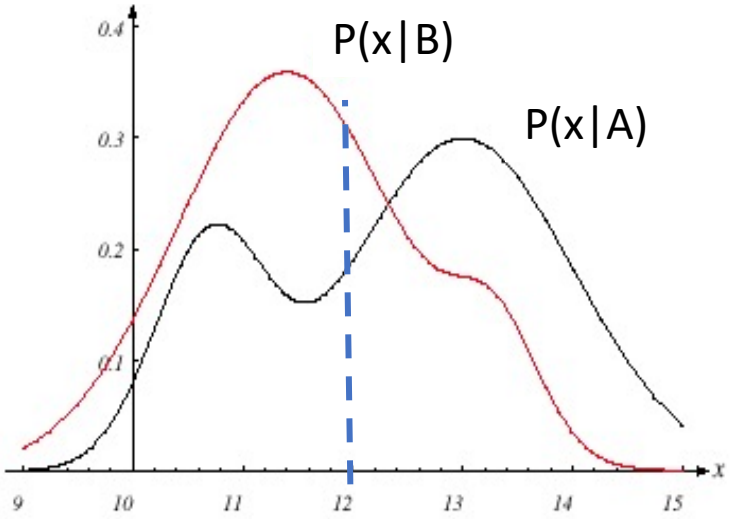
* And also $P(B) = P(B|A)$

Prior probability

- As we said before the simplest way is to ask which class has higher probability in the training set
- What is the probability that you'll pass the exam?
  - We have training data of the previous year result
  - There are 2 classes: 'Pass' or 'Fail'
  - 'Pass' probability is 90%, and 'Fail' is 10%
  - So every one here has 90% to pass the exam
  - This uses the prior probability, and in our context this is the class distribution in the training set

Likelihood

- The likelihood is the class conditional information – the probability of an instance, given the class
  - for an instance x, and 2 possible classes A, B:



If x=12, we'll predict B, because P(x|B)>P(x|A)

Likelihood

- If we return to the previous example (fail \ pass) it is like asking:
  - What is the probability that someone learn to the test if he pass the exam
- But, we wanted to know what is the probability to pass \ fail the exam if you'll learn
- So we need a way to go from likelihood to posterior probability

Posterior probability

- Bayes rule:

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

- With this rule we can convert the likelihood to the posterior probability, if we have also the prior probability
- A classifier that classify A if P(A|x)>P(B|x), is a classifier that maximize the posterior probability – MAP
- The classification with MAP depends on both the likelihood and the prior probabilities

Posterior probability

- So if we want to classify according to MAP:
  - We will classify A if

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

$$P(x|A)P(A) > P(x|B)P(B)$$

  - Note that P(x) is removed from both denominators simply because it is the same

Minimum error

- This classification rule is minimizing the error:
  - If we classify B, then the $P(error|x) = P(A|x)$
  - If we classify A, then the $P(error|x) = P(B|x)$

- But, we classify B only if $P(B|x) > P(A|x)$, and therefore the probability of the error is minimal

$$P(error|x) = \min[P(A|x), P(B|x)]$$

Loss

- We can define a loss measure for wrong decision:
  - 0-1 loss (the simplest one):

$$\lambda_{ij} = \lambda\left(Choose\ A_i \middle| A_j\right) = \begin{cases} 1, & if\ i \neq j \\ 0, & if\ i = j \end{cases}$$

Risk (with 0-1 loss)

- After we defined the loss we can define the risk, which is the expected loss (for k classes):

$$R(Choose\ A_i|x) = \sum_{j=1}^{k} \lambda_{ij} P(A_j|x) = \sum_{j \neq i} P(A_j|x) = 1$$

- Classifier that wants to minimize the risk will choose A_i such that:

$$P(A_i|x) > P\left(A_j\middle|x\right) \ \forall j \neq i$$

Bayes for multi-class – Different calculations

- The full posterior probability:

$$g_i(x) = P(A_i|x) = \frac{P(x|A_i)P(A_i)}{\sum_{j=1}^{k} P(x|A_j)\,P(A_j)} \quad = \quad P(x)$$

- Dropping p(x):

$$g_i(x) = P(x|A_i)P(A_i)$$

Bayes for multi-class

- In order to make the classification process more efficient we can use ln():

$$g_i(x) = \ln\big(P(x|A_i)P(A_i)\big) = \ln\big(P(x|A_i)\big) + \ln\big(P(A_i)\big)$$

- It helps reduce multiplication of small number (0-1) and to deal better with normal distribution $e^{f(x)}$
- We can do it because ln() is monotonically increasing

ML-Hypothesis

- If we're going back to the regression task, we can define the hypothesis to be any function $h(x)$: $X \to R$ that belongs to the hypothesis space $h \in H$
- We want to find the most probable hypothesis
- This is a conditional probability problem – find the hypothesis that maximize

$$P(h|D) = P(D|h)P(h)$$

  * posterior probability

- We will assume all $h \in H$ have the same prior probability, and we'll get that the most probable h will be found according to maximum likelihood:

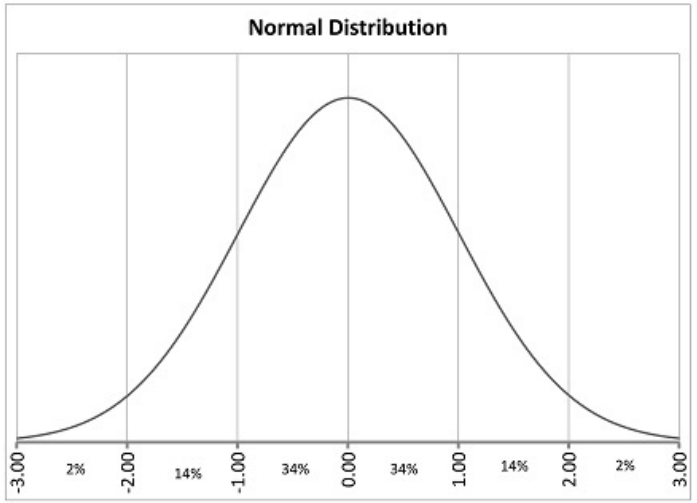
$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)$$

ML-Hypothesis

- Assuming the instances are independent:

$$P(D|h) = \prod_i P(y_i|h)$$

- If the error has normal distribution $e_i \sim N(o, \sigma)$, then we can say that the probability that $h(x_i) = y_i$ is the same as the probability $e_i = 0$ according to the normal distribution of e_i



ML-Hypothesis

- And we get:

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h) = \underset{h \in H}{\operatorname{argmax}} \prod_i P(y_i|h)$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{e_i - 0}{\sigma}\right)^2}$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2}$$

ML-Hypothesis

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i)-y_i}{\sigma}\right)^2}$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \ln\left(\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i)-y_i}{\sigma}\right)^2}\right)$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_i \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left(\frac{h(x_i)-y_i}{\sigma}\right)^2$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_i -\frac{1}{2}\left(\frac{h(x_i)-y_i}{\sigma}\right)^2$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_i -(h(x_i)-y_i)^2$$

$$= \underset{h \in H}{\operatorname{argmin}} \sum_i (h(x_i)-y_i)^2$$

IDC HERZLIYA

Summary so far

- Prior classifier: $P(A) > P(B)$
- ML classifier: $P(x|A) > P(x|B)$ – assuming $P(A) = P(B)$
- MAP classifier:

$$P(A|x) = P(x|A)P(A) > P(x|B)P(B) = P(B|x)$$

* Drooping $P(x)$ from the denominator

How to estimate probabilities

- Parametric estimation
  - If we know \ can guess the distribution type we can estimate the parameters of the distribution
  - Examples
    - Normal Distribution (μ, σ)
    - Poisson (λ)

- Non parametric estimation
  - We don't assume any type of distribution to our data
  - Examples
    - Histogram (=count)
    - Kernel Density Estimation (=smooth histogram)

Parametric

- For each class we will estimate the distribution parameter according to the train dataset
- If we're talking about normal distribution parameters, we need to estimate the mean and the variance:

$$\mu = \frac{1}{m}\sum_{k=1}^{m} x_k$$

$$\sigma^2 = \frac{1}{m}\sum_{k=1}^{m} (x_k - \mu)^2$$

Parametric

- Now, we can estimate the parameter for each likelihood probability, for each class:

$$\mu_i = \frac{1}{|A_i|}\sum_{x \in A_i} x$$

$$\sigma_i{}^2 = \frac{1}{|A_i|}\sum_{x \in A_i} (x - \mu_i)^2$$

- And then classify according to the largest probability given by the normal distribution formula:

$$P(x|A_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

Parametric

- But, this was good only for 1 attribute
- What if we have more than 1?
- In this case each likelihood probability will be estimated according to multivariate normal distribution
- For this we will need mean vector (each dimension will be the mean for some attribute) and the covariance matrix

The covariance matrix

$$\mathbf{S} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Variance

$|\mathbf{S}|$ - is the determinant of the covariance matrix

$\mathbf{S}^{-1}$ - is the inverse matrix of the covariance matrix

IDC HERZLIYA

Parametric

- For each attribute we will find the mean and the variance as before and we will create the mean vector and the covariance matrix
- We will classify according to the multivariate normal distribution:

$$P(\bar{x}|A_i) = \frac{1}{\sqrt{(2\pi)^d |S|}} e^{\left[-\frac{1}{2}(\bar{x}-\overline{\mu_i})^T S^{-1}(\bar{x}-\overline{\mu_i})\right]}$$

Non Parametric

- If we don’t know the type of distribution?
- We need another way to estimate the probabilities $P(x|A_i)$ and $P(A_i)$
- The prior probability $P(A_i)$ can be estimated from the classes frequency in the training set
- But what with the likelihood?
  - You saw in class – Histogram and Interpolation
  - We will see next week another approach.

Dealing with multiple features

- In order to estimate the likelihood for a given instance we need a huge dataset
- If we have d **discrete** attributes, and k **classes** the number of possible terms in the likelihood $P(x_1, x_2, \dots, x_d|A_i)$ is $k \cdot |V_1| \cdot |V_2| \cdots |V_d|$
- We need a way \ assumption to overcome this problem

Naïve Bayes

- If we assume that all attributes are independent **given the class**, we will get:

$$P(x_1, x_2, \dots, x_d | A_i) = \prod_{j=1}^{d} P(x_j | A_i)$$

- And now we can find the MAP:

$$V_{NB} = \underset{i}{\operatorname{argmax}}\, P(A_i) \prod_{j=1}^{d} P(x_j | A_i)$$

- In this assumption we lower the possible size of a **discrete** dataset to

$$k \sum_{j=1}^{d} |V_j|$$

Questions