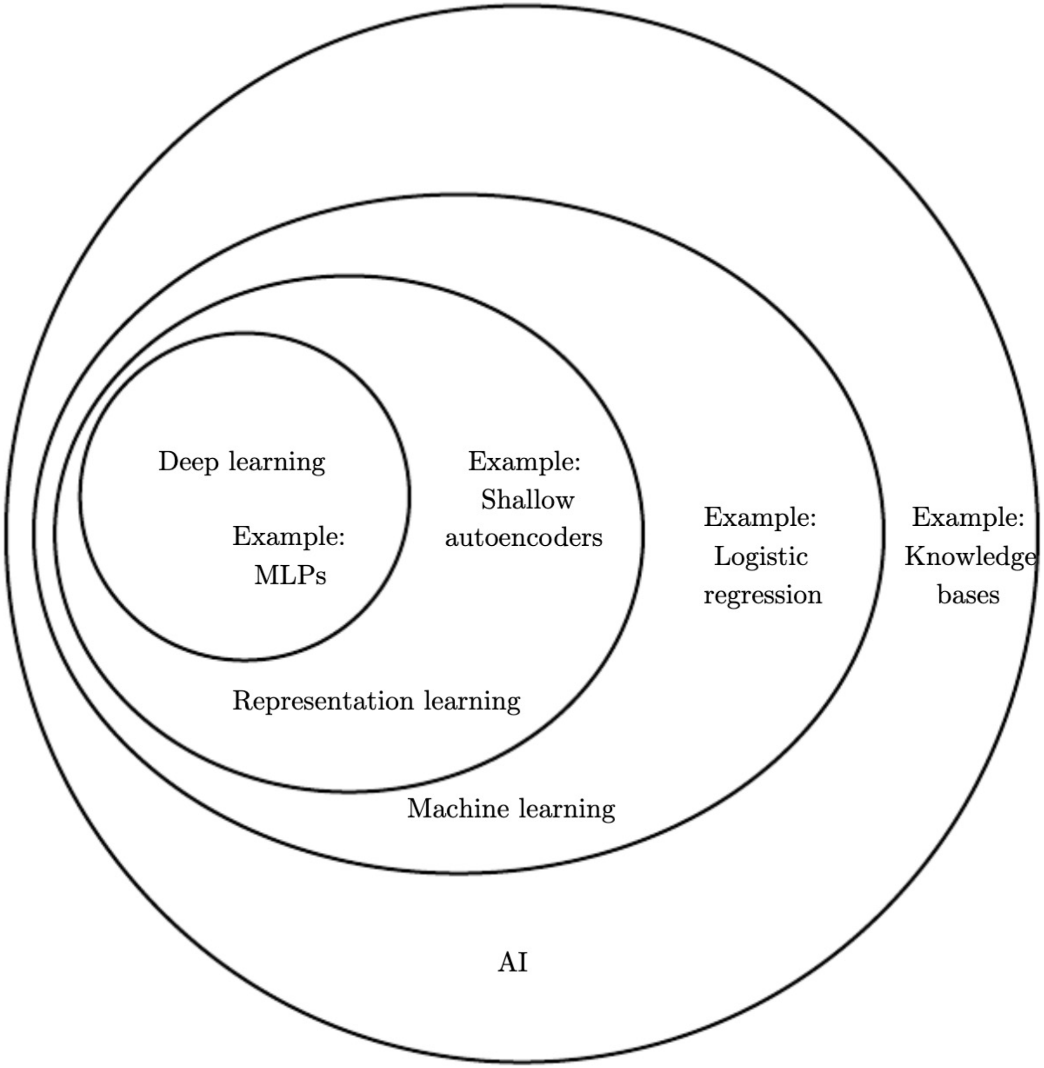
CSE 176 Introduction to Machine Learning

Lecture 2: Linear Algebra, Probability and Statistics

Recap: Different AI systems



Recap: Major Types of machine learning

- Supervised learning: Given pairs of input-output, learn to map the input to output
  - Image classification
  - Speech recognition
  - Regression (continuous output)
- Unsupervised learning: Given unlabeled data, uncover the underlying structure or distribution of the data
  - Clustering
  - Dimensionality reduction
- Reinforcement learning: training an agent to make decisions within an environment to maximize a cumulative reward
  - Game playing (e.g., AlphaGo)
  - Robot control

Linear Algebra

Linear Algebra Topics

- Scalars, Vectors, Matrices and Tensors
- Multiplying Matrices and Vectors
- Identity and Inverse Matrices
- Linear Dependence and Span
- Norms
- Special kinds of matrices and vectors
- Eigen decomposition
- Singular value decomposition

Scalar, Vector, Matrix, Tensor

- **Scalar**: A single number (real-valued or integer)
- **Vector**: An array of numbers arranged in order

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- **Matrix**: 2D Array of numbers

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- **Tensor**: Sometimes need an array with more than two axes
  - E.g., an RGB color image has three axes

Types of matrices

$$\begin{bmatrix} 1 & 3 \\ -4 & 7 \end{bmatrix}$$

Square matrix
2×2

$$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$$

Rectangular matrix
3×2

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Zero matrix
3×5

$$\begin{bmatrix} 1 & 9 & -3 & 0 \end{bmatrix}$$

Row matrix
1×4

$$\begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix}$$

Column matrix
3×1

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Identity matrix
3×3

Matrix times matrix

– If $\boldsymbol{A}$ is of shape mxn and B is of shape nxp then *matrix product* $\boldsymbol{C}$ is of shape mxp

$$C = AB \Rightarrow C_{i,j} = \sum_k A_{i,k} B_{k,j}$$

– Note that the standard product of two matrices is not just the product of two individual elements

- Such a product does exist and is called the element-wise product or the Hadamard product $\boldsymbol{A} \odot \boldsymbol{B}$

Matrix times vector: Linear transformation

- $\boldsymbol{Ax=b}$
  - where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^n$
  - More explicitly

$$
\begin{aligned}
A_{11}x_1 + A_{12}x_2 + \ldots. + A_{1n}x_n &= b_1 \\
A_{21}x_1 + A_{22}x_2 + \ldots. + A_{2n}x_n &= b_2 \\
\\
A_{n1}x_1 + A_{m2}x_2 + \ldots. + A_{n,n}x_n &= b_n
\end{aligned}
$$

n equations in
n unknowns

$$
A = \begin{bmatrix} A_{1,1} & \cdot & A_{1,n} \\ A_{n,1} & \cdot & A_{nn} \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_n \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_n \end{bmatrix}
$$

n x n $\qquad$ n x 1 $\qquad$ n x 1

Can view A as a *linear transformation* of vector $\boldsymbol{x}$ to vector $\boldsymbol{b}$

- Sometimes we wish to solve for the unknowns $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ when A and $\boldsymbol{b}$ provide constraints

UCMERCED

Slide from S. Srihari

Matrix inverse

- Inverse of square matrix A defined as $A^{-1}A = I_n$
- We can now solve $A\boldsymbol{x}=\boldsymbol{b}$ as follows:

$$A\boldsymbol{x} = \boldsymbol{b}$$
$$A^{-1}A\boldsymbol{x} = A^{-1}\boldsymbol{b}$$
$$I_n\boldsymbol{x} = A^{-1}\boldsymbol{b}$$
$$\boldsymbol{x} = A^{-1}\boldsymbol{b}$$

- This depends on being able to find A^{-1}

Matrix inverse

- For a 2x2 matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- The inverse is :

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

- Quiz: find the inverse of AB^T

$$A = \begin{bmatrix} 3 & -1 & 1 \\ 2 & 0 & 2 \end{bmatrix}, B = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

- Answer:

$$AB^T = \begin{bmatrix} 3 & -1 & 1 \\ 2 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 \\ 2 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 6 & 2 \end{bmatrix}$$

$$(AB^T)^{-1} = \frac{1}{5 \cdot 2 - 0 \cdot 6} \cdot \begin{bmatrix} 2 & 0 \\ -6 & 5 \end{bmatrix} = \frac{1}{10} \cdot \begin{bmatrix} 2 & 0 \\ -6 & 5 \end{bmatrix} = \begin{bmatrix} 0.2 & 0 \\ -0.6 & 0.5 \end{bmatrix}$$

Norms

- Used for measuring the size of a vector
- Norms map vectors to non-negative values
- Norm of vector $\boldsymbol{x}=[x_1,..,x_n]^T$ is distance from origin to $\boldsymbol{x}$
  - It is any function f that satisfies:

$$f(\boldsymbol{x})=0 \Rightarrow \boldsymbol{x}=0$$

$$f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x})+f(\boldsymbol{y}) \quad \text{Triangle Inequality}$$

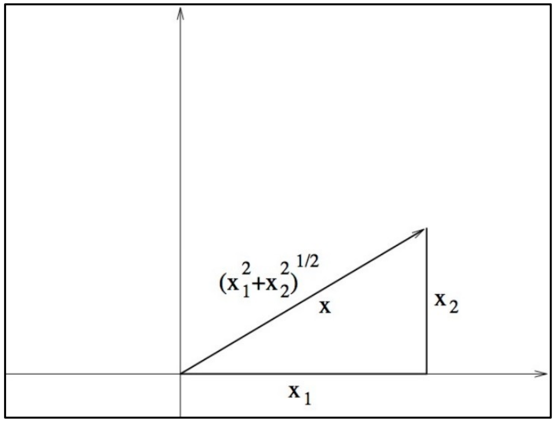
$$\forall \alpha \in R \quad f(\alpha \boldsymbol{x})=|\alpha| f(\boldsymbol{x})$$

L^p Norm

- Definition:

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

  - L^2 Norm
    - Called Euclidean norm
      - Simply the Euclidean distance between the origin and the point $\boldsymbol{x}$
      - written simply as $\|\boldsymbol{x}\|$
      - Squared Euclidean norm is same as $\boldsymbol{x}^T\boldsymbol{x}$



  - L^1 Norm
    - Sum of absolute value for each x_i
  - L^∞ Norm

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|$$

    - Called max norm

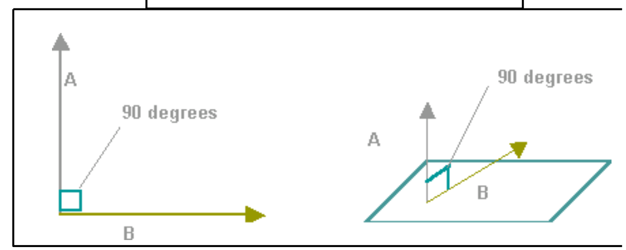
Special kind of vectors

- Unit Vector
  - A vector with unit norm

$$\|\boldsymbol{x}\|_2 = 1$$

- Orthogonal Vectors

$$\begin{bmatrix} 2 \\ -3 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 \\ -2 \\ 7 \end{bmatrix}$$



  - A vector $\boldsymbol{x}$ and a vector $\boldsymbol{y}$ are orthogonal to each other if $\boldsymbol{x}^T\boldsymbol{y}=0$
  - Orthonormal Vectors
    - Vectors are orthogonal & have unit norm
    - Orthogonal Matrix
      - A square matrix whose rows are mutually
      - orthonormal: $A^TA=AA^T=I$

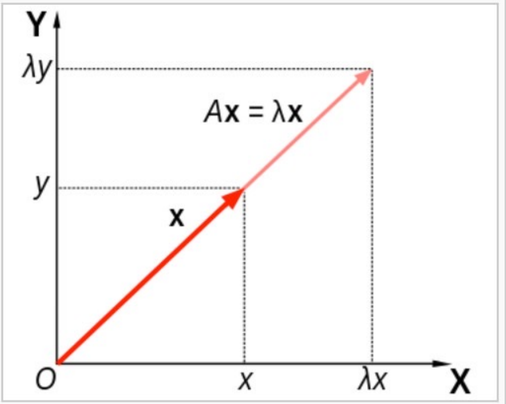
        $A^{-1}=A^T$

Eigenvector

- An eigenvector of a square matrix ***A*** is a non-zero vector ***v*** such that multiplication by ***A*** only changes the scale of ***v***

$$\boldsymbol{Av}=\lambda\boldsymbol{v}$$

  - The scalar λ is known as eigenvalue
- If ***v*** is an eigenvector of ***A***, so is any rescaled vector ***sv***. Moreover ***sv*** still has the same eigen value. Thus look for a unit eigenvector



Matrix *A* acts by stretching the vector *x*, not changing its direction, so *x* is an eigenvector of *A*.

Wikipedia

Slide from S. Srihari

Eigendecomposition

- Suppose that matrix A has n linearly independent eigenvectors $\{\boldsymbol{v}^{(1)},..,\boldsymbol{v}^{(n)}\}$ with eigenvalues $\{\lambda_1,..,\lambda_n\}$
- Concatenate eigenvectors to form matrix V
- Concatenate eigenvalues to form vector $\lambda=[\lambda_1,..,\lambda_n]$
- Eigendecomposition of A is given by

$$A=V\mathrm{diag}(\lambda)V^{-1}$$

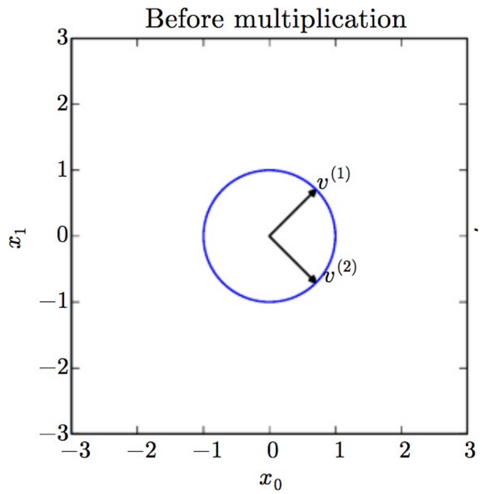
Slide from S. Srihari

Effect of eigenvalue and eigenvector

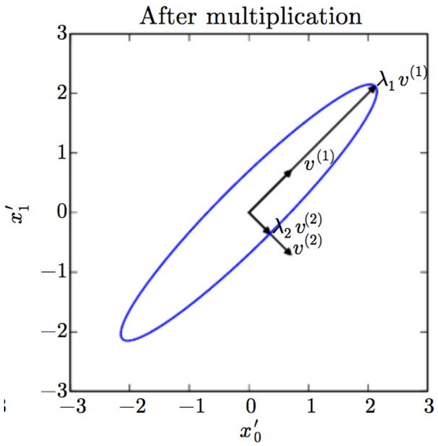
- Example of 2×2 matrix
- Matrix A with two orthonormal eigenvectors
  - – $v^{(1)}$ with eigenvalue λ_1, $v^{(2)}$ with eigenvalue λ_2

Plot of unit vectors $\boldsymbol{u} \in \mathbb{U}^2$ (circle)

Plot of vectors $A\boldsymbol{u}$ (ellipse)



with two variables x_1 and x_2



Positive Semidefinite Matrix (PSD)

- A matrix whose eigenvalues are all positive is called *positive definite*
  - Positive or zero is called *positive semidefinite*
- If eigen values are all negative it is *negative definite*
  - Positive definite matrices guarantee that $\boldsymbol{x}^T A \boldsymbol{x} \geq 0$

Slide from S. Srihari

Singular Value Decomposition (SVD)

- Eigendecomposition has form: $A = V\text{diag}(\lambda)V^{-1}$
  - If A is not square, eigendecomposition is undefined
- SVD is a decomposition of the form $A = UDV^T$
- SVD is more general than eigendecomposition
  - Used with any matrix rather than symmetric ones
  - Every real matrix has a SVD
    - Same is not true of eigen decomposition

Slide from S. Srihari

Probability and Statistics

Random Variable

- Variable that can take different values randomly
- Scalar random variable denoted x
- Vector random variable is denoted in bold as
- Values of r.v.s denoted in italics x or $\boldsymbol{x}$
  - Values denoted as $\mathrm{Val}(\mathrm{x})=\{x_1,x_2\}$
- Random variable must has a probability distribution to specify how likely the states are
- Random variables can be discrete or continuous
  - Discrete values need not be integers, can be named states
  - Continuous random variable is associated with a real value

Slide from S. Srihari

Probability Distribution

- A probability distribution is a description of how likely a random variable or a set of random variables is to take each of its possible states
- The way to describe the distribution depends on whether it is discrete or continuous

Continuous Variables and PDFs

- When working with continuous variables, we describe probability distributions using probability density functions
- To be a pdf p must satisfy:

- The domain of p must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

Slide from S. Srihari

Marginal distribution

- Sometimes we know the joint distribution of several variables
- And we want to know the distribution over some of them
- It can be computed using

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y)$$

$$p(x) = \int p(x, y) dy.$$

Slide from S. Srihari

Conditional probability

- We are often interested in the probability of an event given that some other event has happened
- This is called conditional probability
- It can be computed using

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$

Chain rule of conditional probability

- Any probability distribution over many variables can be decomposed into conditional distributions over only one variable

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

- An example with three variables

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \end{aligned}$$

Slide from S. Srihari

Independence and conditional independence

- Independence: $\boxed{\mathrm{x} \perp \mathrm{y}}$
  - Two variables x and y are independent if their probability distribution can be expressed as a product of two factors, one involving only x and the other involving only y

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x}=x, \mathrm{y}=y) = p(\mathrm{x}=x)p(\mathrm{y}=y)$$

- Conditional Independence: $\boxed{\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}}$
  - Two variables x and y are independent given variable z, if the conditional probability distribution over x. and y factorizes in this way for every z

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z) = p(\mathrm{x}=x \mid \mathrm{z}=z)p(\mathrm{y}=y \mid \mathrm{z}=z)$$

Common probability distribution

- Several simple probability distributions are useful in may contexts in machine learning
  - Bernoulli over a single binary random variable
  - Multinoulli distribution over a variable with k states
  - Gaussian distribution
  - Mixture distribution

Mixture of Distribution

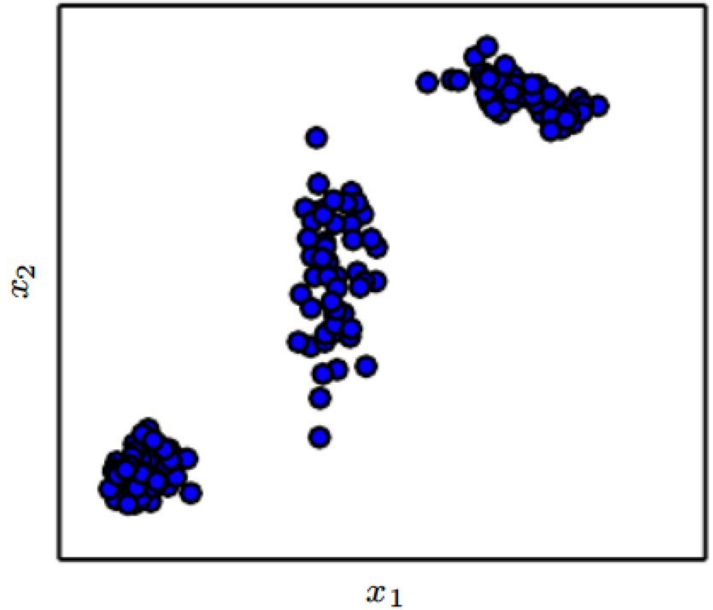
- A mixture distribution is made up of several component distributions
- On each trial, the choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution:

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

  – where $P(\mathrm{c})$ is a multinoulli distribution

Gaussian mixture model

- Components $p(\mathrm{x}|\mathrm{c}{=}i)$ are Gaussian
- Each component has a separately parameterized mean $\mu^{(i)}$ and covariance $\Sigma^{(i)}$
- Any smooth density can be approximated with enough components
- Samples from a GMM:
  - 3 components



Slide from S. Srihari

Quiz

- A random variable, X, has the probability distribution table as shown.

x	-2	-1	0	1	2
$P(X = x)$			0.4	0.1	0.1

Assume that $P(X = -2) = P(X = -1)$. Compute the expectation and variance of X.

Bayes's rule

- **Bayes' theorem** (alternatively **Bayes' law** or **Bayes' rule**), named after Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
- For example, if the risk of health problems is known to increase with age, Bayes' theorem allows the risk to an individual of a known age to be assessed more accurately by conditioning it relative to their age.

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

Quiz

Suppose that $P(A \cap B) = 0.4$ and $P(B) = 0.9$. Find $P(A|B)$.

Solution: $\frac{4}{9}$

From the definition of conditional probability:

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.4}{0.9} = \frac{4}{9} = 0.\bar{4}$

Quiz

- A motor insurance company insures drivers in age group A, B and C. 40% of the customers are in group A, 25% are in B, and 35% are in group C. The company's record shows that each year, 2% of customers in age group A, 1% in group B and 1.5% in group C made a claim. Given that a driver made a claim, what is the probability that the driver is from age group C?