1

一个向量表示一组有序排列的数。通过次序中的索引，我们可以确定每个单独的数。将向量作为空间中的点的几何解释允许我们将ML中的一组输入示例的训练视为空间中点的集合

2

基于距离的相似度计算：
Euclidean Distance，Mahattan Distance，Chebyshev Distance,Minkowski Distance，Mahalanobis Diostance，Lance Williams Distance
基于夹角余弦的相似度计算：
Cosine,Tanimoto Coefficient
前者一般用在图像学领域，后者一般在nlp中使用广泛。基于距离的度量能很好描述特征空间元素的邻域性质，而夹角余弦类能很好描述词句的相似性。

3

超平面是一个子空间，其维数比其环境空间的维数小。在d维向量空间中，超平面具有d-1维，并将空间划分为两个半空间。超平面是高维空间中平面概念的推广
在ML中，超平面是用于线性分类的决策边界， 落在超平面两侧的数据点归因于不同的类

4

矩阵是具有相同特征和纬度的对象的集合，表现为一张二维数据表。其意义是一个对象表示为矩阵中的一行，一个特征表示为矩阵中的一列，每个特征都有数值型的取值。
矩阵在数据分析中作为数据的集合，比如一个batch加入训练。

5

如果存在系数 $a_1, a_2, \ldots, a_k$ 不等于零, 则向量 $v_1, v_2, \ldots, v_k$ 的集合是线性相关的, 因 $\sum_{i=1}^{k} a_i v_i = 0$
如果没有线性依赖性 则向量是线性独立

6

对于nxm矩阵，矩阵的秩是线性独立列的最大数量
满秩代表矩阵中各向量线性无关，秩代表了线性无关列的数量

7

- 矩阵的伪逆，也称为摩尔-彭罗斯伪逆 (Moore-Penrose pseudo-inyerste)。对于不是正方形的矩阵, 逆矩阵不存在

- 因此, 使用伪逆如果 $m > n$, 则伪逆是 $\mathbf{A}^{+} = \left(\mathrm{A}^{\mathrm{T}} \mathrm{A}\right)^{-1} \mathrm{A}^{\mathrm{T}}$ 和 $\mathbf{A}^{+} \mathrm{A} = 1$如果 $m < n$, 则伪逆是 $\mathbf{A}^{\dagger} = \mathbf{A}^{T}\left(\mathrm{AA}^{T}\right)^{-1}$ 和 $\mathbf{A}^{\dagger} \mathbf{A} = \mathbf{I}$

8

n维流形被定义为一个拓扑空间，其性质是每个点都有一个同胚于n维欧几里得空间的邻域。非正式地说, 欧几里得空间是局部平滑的它没有孔, 边缘或其他突然的变化, 也没有相交的邻域。虽然流形在大尺度上可以具有非常复杂的结构, 但在小尺度上与欧几里得空间的相似性允许应用标准的数学概念
流形学习（manifold learning）假设数据在高维空间的分布位于某一更低维的流形上，基于这个假设来进行数据的分析。对于降维，要保证降维之后的数据同样满足与高维空间流形有关的几何约束关系。（比如LLE）

9

我们将一个矩阵A的转置乘以A，并对A^TA求特征值，则有下面的形式：

$$(A^TA)V = \lambda V$$

这里V就是上面的右奇异向量，另外还有：

$$\sigma_i = \sqrt{\lambda_i}, u_i = \frac{1}{\sigma_i}AV$$

这里的σ就是奇异值，u就是上面说的左奇异向量。
奇异值σ跟特征值类似，在矩阵$\sum$中也是从大到小排列，而且σ的减少特别的快，在很多情况下，前10%甚至1%的奇异值的和就占了全部的奇异值之和的99%以上了。也就是说，我们也可以用前r（r远小于m、n）个的奇异值来近似描述矩阵，即部分奇异值分解：

$$A_{m\times n} \approx U_{m\times r}\sum_{r\times r}V^T_{r\times n}$$

右边的三个矩阵相乘的结果将会是一个接近于A的矩阵，在这儿，r越接近于n，则相乘的结果越接近于A。

10

矩阵的F范数：矩阵的各个元素平方之和再开平方根，它通常也叫做矩阵的L2范数，它的优点在于它是一个凸函数，可以求导求解，易于计算

$$\|A\|_F = \sqrt{(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2)}$$

可以优化loss，比较真实的矩阵和估计的矩阵值之间的误差，类似于向量的欧氏距离

11

微分在几何上是导数的斜率，可以近似用线性变化表征函数在某一点的瞬时变化。

12

我们可以连接多元函数相对于其所有输入变量的偏导数, 以获得函数的梯度向量
多元函数 $f(\mathbf{x})$ 相对于 n 维输入向量 $\mathbf{x}=[x_1, x_2, x_n]$ T 的梯度是 n 个偏导数的向量

$$\nabla_{\mathbf{x}} f(\mathrm{x})=\left[\frac{\partial f(\mathrm{x})}{\partial x_1} \frac{\partial f(\mathrm{x})}{\partial z}, \frac{\partial f(\mathrm{x})}{\partial x_n}\right]^T$$

在 ML 中, 梯度下降算法娇赖于损失函数 $\mathcal{L}$ 相对于模型参数 θ ($\nabla_\theta \mathcal{L}$) 的梯度的相反 方向, 以最小化损失函数
可以通过在相对于输入示例 x ($\nabla_x L$) 的损失 L 的梯度方向上添加扰动来创建对抗性示例, 以最大化损失函数

13

对于目标函数f(x)，如果点x处的值是目标函数在x的整个域上的最小值，则它是全局最小值
如果f(x)在x处的值小于x附近任何其他点的目标函数的值，则它是局部最小值

14

在数学术语中, 函数 f 是凸函数, 如果对于 x_1 的所有点 x_2 和所有 $\lambda \in [0,1]$

$$\lambda f\left(x_1\right)+(1-\lambda) f\left(x_2\right) \geq f\left(\lambda x_1+(1-\lambda) x_2\right)$$

凸优化（convex optimization）是最优化问题中非常重要的一类，也是被研究的很透彻的一类。对于机器学习来说，如果要优化的问题被证明是凸优化问题，则说明此问题可以被比较好的解决

15

相关系数是研究变量之间线性相关程度的量。两个随机变量的相关系数定义为：

$$Corr(x,y)=\frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

相关系数的性质：
1）有界性。相关系数的取值范围是 [-1,1]，可以看成无量纲的协方差。
2）值越接近1，说明两个变量正相关性（线性）越强。越接近-1，说明负相关性越强，当为0时，表示两个变量没有相关性。
熵：
对于遵循概率分布 P 且概率质量函数 $P(X)$ 的离散随机变量 X, 通过熵（或香农熵）的 预期信息量为

$$H(X)=\mathbb{E}_{X \sim P}[I(X)]=-\mathbb{E}_{X \sim P}[\log P(X)]$$

$$H(X)=-\sum_X P(X) \log P(X)$$

如果 X 是一个连续随机变量, 它遵循具有概率密度函数 $P(X)$ 的概率分布 P, 则熵为
$A(X) = -\int_X P(X) \log P(X) dX$
对于连续随机变量, 熵也称为微分熵
KL散度:
库尔巴克-莱布勒 (KL) 散度（或相对熵) 提供了两个概率分布差异的度量.对于同一随机变量 X 上的两个概率分布 $P(X)$ 和 $Q(X)$ K散度为
$D_{KL}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P(X)}{Q(X)}\right]$
对于离散随机变量, 此公式等效于
$D_{KL}(P\|Q) = \sum_X P(X) \log \frac{P(X)}{Q(X)} = -\sum_X P(X) \log \frac{Q(X)}{P(X)}$
交叉熵:
交叉熵与 KL散度密切相关，它被定义为 $\mathrm{H(P)}$ 和KL散度 $\mathrm{D_KZ(P}\boldsymbol{P}\mid \mathrm{Q)}$。

$$CE(P, Q) = H(P) + D_{KL}(P + Q)$$

最大似然:
在 ML 中, 我们希望找到一个参数为 θ 的模型, 最大化数据被分配到正确类 的概率, 即
$\mathrm{argmax}_\theta\, P$ (model | data)