C1-14

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

2016.9.22

第十五集：奇异值分解

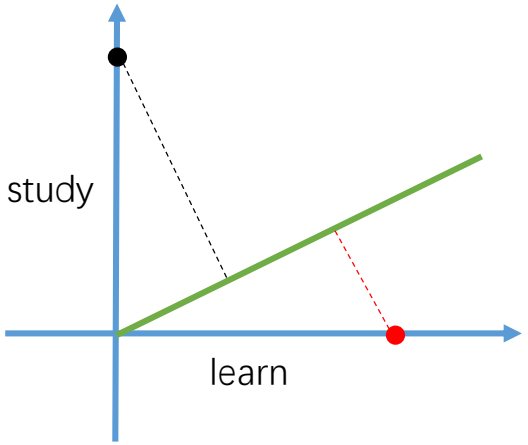
- 隐性语义索引(Latent Semantic Indexing (LSI))
- 奇异值分解(Singularly Valuable Decomposition (SVD))
- 独立成分分析(Independent Components Analysis (ICA))

第十五集：奇异值分解

- 隐性语义索引
  - 主成分分析作为一种高效的降维方法在众多工程领域有着普遍的应用，但是在某些方面如果直接应用会出问题
    - 比如说在图像聚类问题上，对于一幅图像来说假设有100*100个像素点（也就是每个数据样本都拥有100*100个特征）。如果要对图像运用主成分分析进行降维，最终的结果是要计算10000*10000的协方差矩阵的特征值和特征向量，而这个数量显然太大了
    - 与此同时在文本聚类问题上，每一维的特征为邮件中是否出现字典中某一单词。而可以想见的是，每一维的特征的权重应该是不同的。如果我们在预处理过程中强制将每一维特征变成1-方差的，将会无形中忽略了每一维的特征的权重
      - 比如说单词aardvark，在邮件中极少出现。假设10000份邮件中只出现了一次aardvark，因此方差极小。如果在预处理过程中将该维特征的方差变为1，会使得出现的那一次的aardvark的值变得非常大

第十五集：奇异值分解

- 隐性语义索引
  - 而对于文本聚类来说关键在于衡量两篇文档间的相似性
    - 一种容易想见的方法：将文本转换为向量，然后衡量这两个向量之间的夹角
      - 夹角的衡量用cos，计算方法为两个向量的点乘除以各自模的乘积
      - 由于在文本中，每一维的特征表示某个词是否在该文档中出现，因此两个向量的点乘的结果相当于有多少个词在两份文档中同时出现
      - 但如果这样直接衡量两个向量间的夹角，一方面维数高运算量大，另一方面忽略了词相对于文档的信息，比如说第一篇文档中就一个词"study"，第二篇文档中同样就一个词"learn"，乍看上去这两篇文档相似度很高，但是如果用上述方法进行衡量，由于没有同时出现的词而造成相似度为0



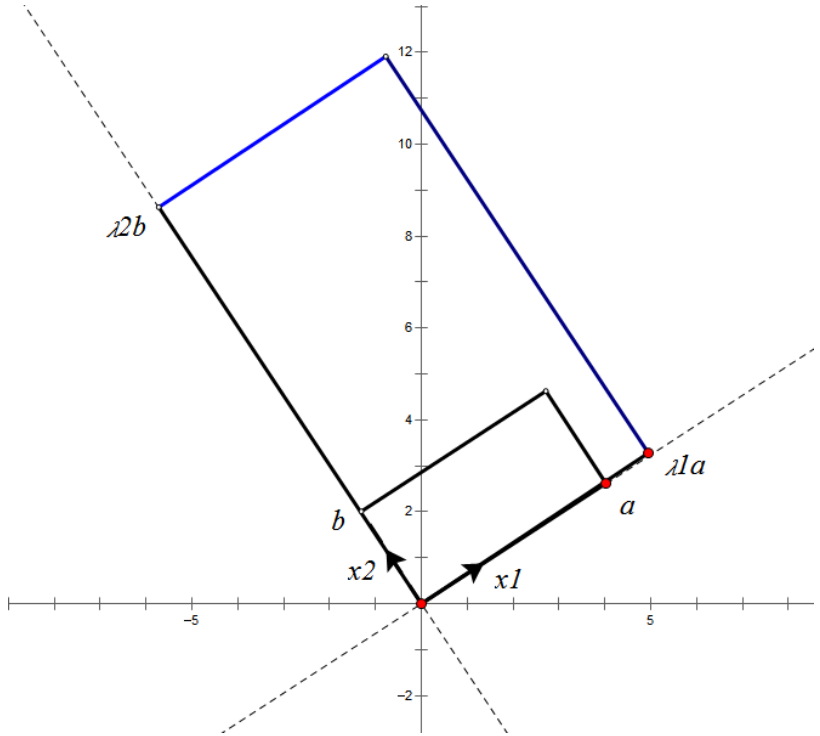
比如此图中黑点表示存在词语"study"而不存在词语"learn"的文档，红点表示存在词语"learn"而不存在词语"study"的文档。由于不同词语对应的特征是相互正交的，因此词语"learn"和词语"study"之间的夹角为90度，两个词语的相似度为0，显然与现实情况不符。而隐性语义索引就是为解决这样的问题，其核心是主成分分析法，分析出一个如图中绿线所示的超平面，然后将样本点投影到这个超平面上，再通过距离的方法判断相似性

第十五集：奇异值分解

- 奇异值分解
  - 下面的问题在于怎样将主成分分析法应用上来
    - 这里面最大的问题是样本个数和维数都非常大，导致协方差矩阵过大无法直接进行存储处理的情况
    - 而这已经无法直接应用之前讲到过的主成分分析法加以解决，而需要采用一种称为奇异值分解的方法
      - 在奇异值分解中，会将m*n的矩阵A分解为m*n的矩阵U，n*n的矩阵D和n*n的矩阵V^T的乘积，其中矩阵D为对角矩阵（除主对角线外其他值均为0）且主对角线上的值被称为矩阵A的奇异值
      - 在介绍奇异值分解前先介绍另外一种更加基本的分解方法，称为特征值分解。在这一方法中能将满秩对称矩阵A分解为m*m的方阵U，m*m的矩阵Λ和m*m的方阵U^T的乘积，其中矩阵Λ为对角矩阵（除主对角线外其他值均为0）且主对角线上的值被称为矩阵A的特征值
        - 实际上之前讲到的主成分分析法也用到了特征值分解的思想，其分解目标是协方差矩阵Σ，将其分解为一系列相互正交的方向向量u所构成的矩阵U和对应特征值λ所构成的矩阵Λ

第十五集：奇异值分解

- 奇异值分解
  - 下面先来介绍一般意义上的特征值分解法



假设 m 阶方阵A为满秩对称矩阵（满秩的意思是矩阵列相互之间线性独立）。由于满秩，因此矩阵A有 m 个不同的特征值$\lambda_1,\lambda_2,\dots,\lambda_m$，而这些特征值对应的特征向量为$u_1,u_2,\dots,u_m$

由于u为特征向量，λ为对应特征值，因此有：

$$\begin{cases} \mathrm{A}u_1=\lambda_1u_1 \\ \mathrm{A}u_2=\lambda_2u_2 \\ \vdots \\ \mathrm{A}u_m=\lambda_mu_m \end{cases} \Leftrightarrow AU=U\Lambda, U=[u_1 \quad \dots \quad u_m], \Lambda=\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix}$$

由于A为对称矩阵，因此其特征向量相互正交，由此可得矩阵U是一个正交矩阵。对于一个正交矩阵来说其逆U^{-1}等于其转置U^T，因此有：

$$A=U\Lambda U^{-1}=U\Lambda U^T$$

而此举对应的功能将在以$u_1,u_2,\dots,u_m$作为基底的空间下将每一维度的点都进行了拉伸变换，拉伸变换的程度由系数$\lambda_1,\lambda_2,\dots,\lambda_m$所控制。假设一点$\mathrm{x}=a_1u_1+a_2u_2+\cdots+a_mu_m$，则有：

$$\mathrm{A}x=U\Lambda U^Tx=U\Lambda\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}\left(a_1u_1+a_2u_2+\cdots+a_mu_m\right)=U\Lambda\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}=U\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

$$=U\begin{bmatrix} \lambda_1a_1 \\ \lambda_2a_2 \\ \vdots \\ \lambda_ma_m \end{bmatrix}=[u_1 \quad \dots \quad u_m]\begin{bmatrix} \lambda_1a_1 \\ \lambda_2a_2 \\ \vdots \\ \lambda_ma_m \end{bmatrix}=a_1\lambda_1u_1+a_2\lambda_2u_2+\cdots+a_m\lambda_mu_m$$

第十五集：奇异值分解

- 奇异值分解
  - 下面将从特征值分解法引出奇异值分解法

在特征值分解法中，我们假定A为满秩对称方阵，并在此要求下找到了一组正交基使得在以$u_1,u_2,\dots,u_m$作为基底的空间下将每一维度的点都进行了拉伸变换。但是实际情况下，很多矩阵A并不保证一定是方阵，而且也不一定满秩。那么在矩阵A是一个$m\times n$的可能并不满秩的矩阵下，是否还能找到这样的一组正交基是奇异值分解的核心问题。

那么一开始我们不妨假设存在这样的一组正交基$v_1,v_2,\dots,v_m$，这组正交基被矩阵A映射后的结果为$\mathrm{A}v_1,\mathrm{A}v_2,\dots,\mathrm{A}v_m$，现在要求它们也互相正交，因此有：

$$\mathrm{A}v_i\cdot\mathrm{A}v_j=\left(\mathrm{A}v_i\right)^T\left(\mathrm{A}v_j\right)=v_i^TA^T\mathrm{A}v_j=0$$

$$v_i\cdot v_j=v_i^Tv_j=0$$

由于要同时满足上面两个约束，因此将下式尽量形式上往上式靠拢，因此引入一个常数λ：

$$0=v_i\cdot v_j=\lambda v_i\cdot v_j=\lambda v_i^Tv_j=v_i^T\lambda v_j=v_i^TA^T\mathrm{A}v_j=0$$

由此可得：

$$\mathrm{A}v_i\cdot\mathrm{A}v_i=v_i^TA^T\mathrm{A}v_i=v_i^T\lambda_iv_i=\lambda_iv_i^Tv_i=\lambda_i$$

其中由于$A^T\mathrm{A}$为对称矩阵（任意矩阵乘以其转置一定对称），因此令$\mathcal{A}=A^T\mathrm{A}$，由于$\mathcal{A}$为对称矩阵，因此有：

$$v_i^T\mathcal{A}v_i=v_i^T\lambda_iv_i\Leftrightarrow\mathcal{A}v_i=\lambda_iv_i$$

由此可知v_i和λ_i分别为$A^T\mathrm{A}$的特征向量和对应的特征值而且可保证特征向量间是相互正交的，由于$\mathrm{A}v_i=\mathrm{A}v_i$，因此有：

$$\mathrm{A}v_i\cdot\mathrm{A}v_i=\left|\mathrm{A}v_i\right|\left|\mathrm{A}v_i\right|=\left|\mathrm{A}v_i\right|^2=\lambda_i$$

由于正交基$v_1,v_2,\dots,v_m$被矩阵A映射后的结果为$\mathrm{A}v_1,\mathrm{A}v_2,\dots,\mathrm{A}v_m$，不妨设$u_1,u_2,\dots,u_m$为对应于$\mathrm{A}v_1,\mathrm{A}v_2,\dots,\mathrm{A}v_m$的单位向量，即：

$$u_i=\frac{\mathrm{A}v_i}{\left|\mathrm{A}v_i\right|}=\frac{1}{\sqrt{\lambda_i}}\mathrm{A}v_i\Leftrightarrow\mathrm{A}v_i=\sqrt{\lambda_i}u_i\Leftrightarrow\mathrm{A}v_i=\sigma_iu_i,\sigma_i=\sqrt{\lambda_i},0\le i\le\mathrm{rank}(\mathrm{A})$$

对比于A为满秩对称方阵下的结论$\mathrm{A}u_i=\lambda_iu_i$发现，在A为普通的$m\times n$矩阵下，虽然并不一定能找到一组正交基使得在以$u_1,u_2,\dots,u_m$作为基底的空间下，将每一维度的点都进行了拉伸变换，但是可以找到一组正交基$v_1,v_2,\dots,v_k$，经过矩阵A的变换后依然得到了一组正交基$u_1,u_2,\dots,u_k$，其中$k=\mathrm{rank}(\mathrm{A})$. 如果此时矩阵A是满秩的，则$k=\mathrm{rank}(\mathrm{A})=n$

$$\mathrm{AV}=\mathrm{A}[v_1\ \ \dots\ \ v_n]=[Av_1\ \ \dots\ \ \mathrm{A}v_k]=[\sigma_1u_1\ \ \dots\ \ \sigma_ku_k]=[u_1\ \ \dots\ \ u_k]\begin{bmatrix}\sigma_1&\cdots&0\\\vdots&\ddots&\vdots\\0&\cdots&\sigma_k\end{bmatrix}=U\Sigma$$

由于矩阵V由$A^T\mathrm{A}$的特征向量得到，而$A^T\mathrm{A}$为对称矩阵，因此矩阵V为对角矩阵，因此上式也等价于如下形式（而这也就是矩阵A的奇异值分解）：

$$\mathrm{AV}=U\Sigma\Leftrightarrow\mathrm{AV}V^{-1}=U\Sigma V^{-1}\Leftrightarrow\mathrm{A}=U\Sigma V^T$$

第十五集：奇异值分解

- 奇异值分解
  - 现在考虑如何通过奇异值分解来应对协方差矩阵过大的问题

对于样本集$x_1, x_2, \ldots, x_m$来说，其协方差矩阵$\Sigma = \sum_{i=1}^{m} \left(x^{(i)} \right) \left(x^{(i)} \right)^T$，由此可以构造矩阵$X$满足下面的形式：

$$X = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix}, X^T = \begin{bmatrix} x^{(1)} & \ldots & x^{(m)} \end{bmatrix}$$

因此有：

$$\Sigma = \begin{bmatrix} x^{(1)} & \ldots & x^{(m)} \end{bmatrix} \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} = X^T X$$

而根据之前所证明的 SVD 方法，我们可以通过对矩阵X进行 SVD 分解（对于 SVD 分解有很多分解的高效算法），从而避免将协方差矩阵Σ表示出来求特征值和特征向量。在 SVD 分解后将奇异值$\sigma_i = \sqrt{\lambda_i}$由大到小依次排列（记得对应调整矩阵$U$和矩阵$V^T$的行列），再选取前r行($\mathrm{r} \leq \mathrm{rank(X)}$)，即可得到对应的$\Sigma = X^T X$的r个最优的特征值$\lambda_i$和特征向量$v_i$，其中特征向量$v_i$即为原样本空间降维到的最优r维空间的正交基

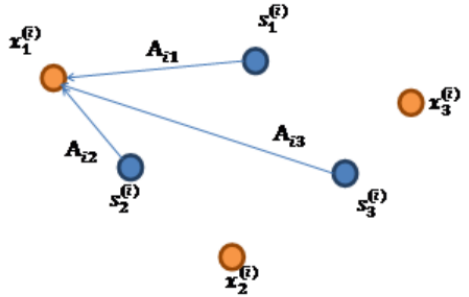
第十五集：奇异值分解

- 奇异值分解
  - 在这最后，对于非监督学习的聚类问题做一个小结

	基于概率密度的方法	非概率密度的方法
样本维数大于样本生成空间	因子分析	主成分分析
样本维数等于样本生成空间	高斯混合模型 朴素贝叶斯混合模型 …	K-均值聚类

第十五集：奇异值分解

- 独立成分分析
  - 独立成分分析和主成分分析都是为了找到表示数据的一组基，但其出发点各不相同
    - 对于主成分分析来说，遇到的问题是样本只能生成样本空间的子空间，解决的关键是找到低维空间的一组基将高维样本映射到低维空间中
    - 而对于独立成分分析来说，遇到的问题是样本由多个独立的成分混杂而成，解决的关键是要找的每个独立成分的基，从而将混杂的样本区别开来
      - 比如说著名的鸡尾酒派对问题，录下的声音中混有多个人的人声，多种乐器的乐器声，还有其他噪声。那怎么将某个人的人声从中区别开来就是独立成分分析中需要解决的问题
      - 现在给出这个问题更加正式的定义：n个独立成分可以看成n个特征，那么可以假设我们最终得到的混合了多种成分的数据x，可以由这n个特征构造出的n维向量s通过线性变换得出，即x=As
        - 其中A为未知的n维方阵，称为混合矩阵(mixing matrix)。在混合矩阵的作用下n个独立成分被杂糅在一起，成为了最终观测到的数据x.我们的目标是已知观测数据x的情况下还原出向量s



如左图所示，图中蓝点表示在鸡尾酒派对上说话的人，黄点表示麦克风。在这个例子中假设三个人在i时刻同时在派对上说话，那么s就是一个3维向量，s_1表示第一个人的语音，s_2表示第二个人的语音，s_3表示第三个人的语音。最终第一个麦克风x_1录下的这一时刻的观测数据为s中数据的线性组合（因为还要考虑说话距离麦克风的距离等因素，因此不是简单的相加）

第十五集：奇异值分解

- 独立成分分析
  - 由于我们是在给定观测数据x的情况下分析其原始成分s
    - 这相当于要找出混合矩阵A的逆W（这也被称为分离矩阵(unmixing matrix)），使得s=Wx，为了记号方便在这里用ω表示分离矩阵W的行向量，即：

$$W = \begin{bmatrix} — & w_1^T & — \\ & \vdots & \\ — & w_n^T & — \end{bmatrix}$$

由此可知向量s的第j个特征s_j（也就是第j个原始成分），可由$\omega_j x$计算得到

    - 这时有个严重的问题，因为我们仅知道观测数据x，而矩阵W和变换后的向量s均未知，那么这会导致无数种矩阵W和向量s的组合情况
      - 举个简单的例子，实数a,b相乘的结果为c，现在给定c问a,b分别是什么？这个问题自然会有无数种答案，因为0.5a*2b=c, a*b=c...由此可以看出如果不对矩阵W和向量s中的某一个加以限制，会出现无数种可能。而且对于矩阵运算来说不仅这种倍乘的情况可以得出同样的结果，其实任意一种初等行变换都可以得到相等的结果（只需要在原矩阵的前提下再乘上一个初等矩阵）

第十五集：奇异值分解

- 独立成分分析
  - 与此同时还有其他可能会造成无数种矩阵W和向量s的组合情况，比如说观测数据不能是旋转对称分布的
    - 比如说标准正态分布就是一种旋转对称分布，对于一个0均值单位方差的标准正态分布来说不论坐标轴怎样旋转，都不影响最终数据样本x的均值和方差，但是却带来了矩阵W和向量s的改变

假设s的维数为 2，并满足标准正态分布$\mathcal{N}(0,I)$，由于高斯分布的任意叠加仍然是高斯分布，因此x = As仍然为高斯分布

$$\mathrm{E}[\mathrm{x}] = E[As] = AE[s] = 0, Var(x) = E\left[xx^T\right] = E\left[Ass^TA^T\right] = AE\left[ss^T\right]A^T = AA^T$$

令R是任意正交阵，不难得到x = ARs同样也是高斯分布，而且均值和方差计算为：

$$\mathrm{E}[\mathrm{x}] = E[ARs] = ARE[s] = 0, Var(x) = E\left[xx^T\right] = E\left[ARss^TR^TA^T\right] = ARE\left[ss^T\right]R^TA^T = ARR^TA^T = AA^T$$

发现此时x的分布与原来相同

第十五集：奇异值分解

- 独立成分分析
  - 经过前面的分析我们发现需要对于s的分布加以限制，但在此之前先看一般情况
    - 假设s中每一维的特征s_i属于概率密度p_s，因此可计算s的联合分布为（在这里认为每一维的特征之间相互独立）
    - 现在的问题是如何通过s的联合分布p(s)得到x的联合分布p(x)

$$p(s) = \prod_{i=1}^{n} p_s(s_i)$$

由于我们已知了$s = Wx$和s的联合概率密度函数$p_s(s)$，现在的问题是如何通过这个求得x的联合概率密度函数$p_x(x)$

对此可能会想当然地认为，令$s = Wx$代入联合概率密度函数$p_s(s)$即可得$p_x(x) = p_s(Wx)$，然而这样并不正确。举个反例，假如$p_s(s)$服从$[0,1]$均匀分布，那么$p_s(s) = 1, s \in [0,1]$.假设矩阵$A = [2]$，可知$p_x(x)$服从$[0,2]$均匀分布，$p_x(x) = 0.5$.但是根据$p_x(x) = p_s(Wx) = p_s(0.5x) = 1$

而实际上，真正的x的联合概率密度函数$p_x(x)$由下式推导出来：

$$F_x(x) = p(X \le x) = p(AS \le x) = p(S \le Wx) = F_s(Wx)$$

$$p_x(x) = F'_x(x) = F'_s(Wx) = p_s(Wx)|W|$$

第十五集：奇异值分解

- 独立成分分析
  - 由此可以将s的联合概率密度函数p_s表示为x的联合概率密度函数p_x，从而可得

$$p(x) = \prod_{i=1}^{n} p_s(w_i^T x) \cdot |W|$$

  - 根据前面的推导发现如果没有对于矩阵W或者S的分布加以限定，最后有无数种求解可能，在这里比较简单的是对于S的分布加以限定
    - 对此一种简单的做法是对于概率分布函数而非概率密度函数加以限定（虽然这两者本质相同，但是概率分布函数到密度函数是微分，比较好求）
      - 根据之前所述不能讲S的分布限定为高斯分布这种旋转对称的分布，而且由于是限定概率分布函数，因此该函数的必须在定义域内递增且值域在[0,1]区间内
      - 与此同时还要微分比较好算，对此发现Sigmoid函数非常合适

$$g(s) = \mathrm{F}_s(\mathrm{s}) = \frac{1}{1+e^{-s}} \Leftrightarrow g'(s) = \mathrm{p}_s(\mathrm{s}) = \frac{e^{-s}}{\left(1+e^{-s}\right)^2} = \frac{\frac{1}{e^s}}{1+\frac{1}{e^{2s}}+\frac{2}{e^s}} = \frac{e^s}{\left(1+e^s\right)^2}$$

第十五集：奇异值分解

- 独立成分分析
  - 由此通过得到似然方程即可得最终矩阵W的解

$$\ell(\mathrm{W}) = \log\left[\prod_{i=1}^{m} \mathrm{p}_x\left(\mathrm{x}^{(i)}\right)\right]$$

$$= \log\left[\prod_{i=1}^{m}\left(\mathrm{p}_s\left(\mathrm{Wx}^{(i)}\right)|\mathrm{W}|\right)\right] = \log\left[\prod_{i=1}^{m}\left(\frac{e^{\mathrm{Wx}^{(i)}}}{\left(1+e^{\mathrm{Wx}^{(i)}}\right)^2}|\mathrm{W}|\right)\right]$$

$$= \sum_{i=1}^{m}\left(\log\frac{e^{\mathrm{Wx}^{(i)}}}{\left(1+e^{\mathrm{Wx}^{(i)}}\right)^2} + \log|\mathrm{W}|\right) = \sum_{i=1}^{m}\left(\mathrm{Wx}^{(i)} - 2\log\left(1+e^{\mathrm{Wx}^{(i)}}\right) + \log|\mathrm{W}|\right)$$

$$W := W + \alpha\left(\begin{bmatrix} 1-2g(w_1^T x^{(i)}) \\ 1-2g(w_2^T x^{(i)}) \\ \vdots \\ 1-2g(w_n^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^T)^{-1}\right)$$

$$\nabla_W \ell(\mathrm{W}) = \nabla_W \sum_{i=1}^{m}\left(\mathrm{Wx}^{(i)} - 2\log\left(1+e^{\mathrm{Wx}^{(i)}}\right) + \log|\mathrm{W}|\right)$$

$$= \sum_{i=1}^{m}\left(\mathrm{x}^{(i)} - 2\frac{xe^{\mathrm{Wx}^{(i)}}}{1+e^{\mathrm{Wx}^{(i)}}} + \frac{\left(adj(\mathrm{W})\right)^T}{|\mathrm{W}|}\right) = \sum_{i=1}^{m}\left(\frac{1-e^{\mathrm{Wx}^{(i)}}}{1+e^{\mathrm{Wx}^{(i)}}}\mathrm{x}^{(i)} + \frac{|\mathrm{W}|\left(\mathrm{W}^{-1}\right)^T}{|\mathrm{W}|}\right)$$

$$= \sum_{i=1}^{m}\left(\left(1-\frac{2}{1+e^{-\mathrm{Wx}^{(i)}}}\right)\mathrm{x}^{(i)} + \left(\mathrm{W}^{-1}\right)^T\right) = \sum_{i=1}^{m}\left(\left(1-2g(\mathrm{Wx}^{(i)})\right)\mathrm{x}^{(i)} + \left(\mathrm{W}^{-1}\right)^T\right)$$

文档结尾

由此通过如左所示的梯度下降法即可完成对于矩阵W的求解，向量s可直接由Wx得到