2016－2017 学年第一学期期末试卷

《 机器学习 答案》（A 卷）

图模型提供了一种描述框架：结点表示随机变量（集合），边表示变量之间的依赖关系。

一. 假设在织物疵点检测中正品（ w_1 ）和次品（ w_2 ）两类的先验概率分别为 $p(w_1)=0.8$， $p(w_2)=0.2$，现有一待检测织物，其观察值为 x，从类条件概率密度分布上曲线上查得 $p(x|w_1)=0.25$，$p(x|w_2)=0.6$，并且已知决策表如表 1 所示。试对该织物 x 用以下两种方法进行分类：（10 分）

（1）基于最小错误率的贝叶斯决策；

（2）基于最小风险的贝叶斯决策。

表 1 决策表

决策	状态	
	w_1	w_2
α_1	0	6
α_2	1	0

答：（1）利用贝叶斯公式，分别计算出 w_1 和 w_2 的后验概率：

$$p(w_1|x)=\frac{p(x|w_1)\,p(w_1)}{\sum_{j=1}^{2}p(x|w_j)\,p(w_{j1})}=\frac{0.25\times0.8}{0.25\times0.8+0.6\times0.2}=0.625$$

$$p(w_2|x)=1-p(w_1|x)=0.375$$

根据贝叶斯决策规则，$p(w_1|x)=0.625>p(w_2|x)=0.375$，所以把 x 归为正品。

（2）根据条件和上面算出的后验概率，计算出条件风险：

$$R(\alpha_1|x)=\sum_{j=1}^{2}\lambda_{1j}p(w_j|x)=\lambda_{11}p(w_1|x)+\lambda_{12}p(w_2|x)=3.125$$

$$R(\alpha_2|x)=\sum_{j=1}^{2}\lambda_{2j}p(w_j|x)=\lambda_{21}p(w_1|x)+\lambda_{22}p(w_2|x)=1.75$$

由于 $R(\alpha_1|x)>R(\alpha_2|x)$，即决策为 w_2 的条件风险小于决策为 w_1 的条件风险，因此采取决策行动 α_2，即判断待检测的织布 x 为 w_2 类--次品。

二. 请叙述预剪枝和后剪枝的技术原理及优缺点。（10 分）

答：决策树是一种树形结构，由结点和有向边组成。决策树对训练数据有很好的分类能力，但对未知的测试数据未必有好的分类能力，泛化能力弱，即可能

发生过拟合现象。针对过拟合问题，剪枝是主要手段。

预剪枝策略(Pre-pruning)：决策树生成过程中，对每个结点在划分前进行估计，若划分不能带来决策树泛化性能提升，则停止划分并将该节点设为叶结点。

优势：“剪掉”很多没必要展开的分支，降低了过拟合风险，并且显著减少了决策树的训练时间开销和测试时间开销.

劣势：有些分支的当前划分有可能不能提高甚至降低泛化性能，但后续划分有可能提高泛化性能；预剪枝禁止这些后续分支的展开，可能会导致欠拟合。

后剪枝策略(Post-pruning)：先利用训练集生成决策树，自底向上对非叶结点进行考察，若将该结点对应子树替换为叶结点能带来泛化性能提升，则将该子树替换为叶结点。

优势：测试了所有分支，比预剪枝决策树保留了更多分支，降低了欠拟合的风险，泛化性能一般优于预剪枝决策树。

劣势：后剪枝过程在生成完全决策树后再进行，且要自底向上对所有非叶节点逐一评估;因此，决策树的训练时间开销要高于未剪枝决策树和预剪枝决策树。

三. 以 K 均值（K-means）聚类为例，描述期望最大化算法（EM 算法）的步骤和基本原理。（10 分）

对于样本 $\mathbf{x}_n$，定义一个聚类标注 r_n，即如果 $\mathbf{x}_n$ 属于第k个聚类，则

$$r_{nk} = 1, \text{and } r_{nj} = 0 \text{ for } j \neq k$$

准则函数：

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

$\boldsymbol{\mu}_k$

r_{nk}

两步走策略

– 第一步：初始化μ_k，按照最优化准则产生r_{nk}

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \parallel \mathbf{X}_n - \boldsymbol{\mu}_k \parallel^2$$

$$\Rightarrow \quad r_{nk} = \begin{cases} 0 \text{ if } k = \arg\min_j ||\mathbf{X}_n - \boldsymbol{\mu}_j||^2 \\ 1 \text{ otherwise} \end{cases}$$

– 第二步：初始化r_{nk}，按照最优化准则产生μ_k

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \parallel \mathbf{X}_n - \boldsymbol{\mu}_k \parallel^2$$

$$\Rightarrow \quad 2\sum_{n=1}^{N} r_{nk}(\mathbf{X}_n - \boldsymbol{\mu}_k) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{X}_n}{\sum_n r_{nk}}$$

两步走策略

–循环迭代：根据产生的μ_k ，按照最优准则产生γ_{nk}

迭代γ_{nk} ----Expectation

迭代μ_k -----Maximization

四. 简述集成学习（Ensemble Learning）的概念，并选择一种串行化方法和并行化方法解释其原理。（10 分）

答：集成学习(EnsembleLearning)是通过构建并结合多个分类器完成学习任务，也称为多分类器系统(Multi-Classifier System)、基于委员会的学(Committee based Learning)等。通过将多个学习器进行整合，常可获得比单一学习器显著优越的泛化性能，这对弱分类器尤为明显。

串行化方法可将弱学习器提升为强学习器。个体学习器存在强依赖关系；串行生成；每次调整训练数据的样本分布。以Boosting为例，先从初始数据集训练出一个基学习器，再根据其对训练样本分布进行调整，使先前做错的样本在后续受到更多关注，然后基于调整后的样本分布训练下一个基学习器；重复进行直至基学习器数目达到预先指定值。最终将这些基学习器加权结合。

并行化方法：Bagging算法利用自助法采样可构造T个含m个训练样本的采样

集，基于每个采样集训练出一个基学习器，再将它们进行结合(在对预测输出结合时，通常对分类任务使用简单投票法，对回归任务使用简单平均法)。

五. 简述支持向量机（SVM）与核技术原理。（10 分）

答：支持向量机是基于统计学习理论发展起来的一种新的机器学习的方法。SVM 从线性可分情况下的最优分类面发展而来。最优分类面就是要求分类线不但能将两类正确分开(训练错误率为 0)，且使分类间隔最大。SVM 考虑寻找一个满足分类要求的超平面，并且使训练集中的点距离分类面尽可能的远，也就是寻找一个分类面使它两侧的空白区域(Margin)最大。

非线性支持向量机利用一个固定的非线性映射将数据映射到特征空间学习的线性分类器等价于基于原始数据学习的非线性分类器。核函数在特征空间中直接计算数据映射后的内积就像在原始输入数据的函数中计算一样，大大简化了计算过程。常用的核函数有：线性核，多项式核，高斯核以及Sigmoid核。

六.（1）列出图 1(a)中所有的极大团；（4 分）

（2）令 $X=\{x_1,x_2,\ldots,x_7\}$，分别写出图 1(a)和(b)中变量的联合概率分布（用$\psi_Q(X_Q)$表示定义在团 X_Q 上的势函数）。（6 分）

(a)马尔可夫随机场　　(b)贝叶斯网络

图 1

(1)答： 对于图中结点的一个子集，若其中任意两结点间都有边连接，则称该结点子集为一个“团”（clique）。若一个团中加入另外任何一个结点都不再形成团，则称该团为“极大团”（maximal clique)

•图中{x1,x2},{ x1,x3},{ x2,x4},{ x2,x5},{ x2,x6},{ x3,x5},{ x5,x6}和{ x2,x5,x6}都是团，除了{ x2,x5},{ x2,x6}和{ x5,x6}之外都是极大团。每个结点至少出现在一个极大团中；多个变量之间的连续分布可基于团分解为多个因子的乘积。

（2）

（a）马尔科夫随机场的联合概率定义为：$P(\mathrm{x})=\frac{1}{Z}\prod_{Q\in C}\varphi(\mathrm{x}_Q)$，其中Z为规范化因子

$\mathbf{x}=\{x_1,x_2,\cdots,x_6\}$，联合概率分布$P(\mathbf{x})$定义为：

$$P(\mathbf{x})=\frac{1}{Z}\varphi_{12}(x_1,x_2)\varphi_{13}(x_1,x_3)\varphi_{24}(x_2,x_4)\varphi_{35}(x_3,x_5)\varphi_{256}(x_2,x_5,x_6)$$

其中，势函数$\varphi_{256}(x_2,x_5,x_6)$定义在极大团{ x2, x5, x6}上，由于它的存在，不再需为团{ x2, x5}，{ x2, x6}和{ x5, x6}构建势函数。

（b）贝叶斯网络的联合概率定义为：$P(\mathrm{x})=\prod_{i\in I}p(x_i\mid x_{pa(i)})$

$$P(\mathbf{x})=p(x_1)\,p(x_2\mid x_1)\,p(x_3\mid x_1)\,p(x_4\mid x_2)\,p(x_5\mid x_2,x_3)\,p(x_6\mid x_2,x_5)$$

请从下面三题中选做两题。

七. 将D维数据集$\{x_n\}$，$n=1,2,\cdots,N$降为M维（$M<D$），请从最大方差思想或最小均方误差思想角度推导主成分分析（Principal Component Analysis）的计算过程。（20 分，选做）

答：

- **最大方差思想**

使用较少的数据维度保留住较多的原数据特性

将D维数据集 $\{\mathbf{x}_n\},n=1,2,...,N$ **降为M维，**$M<D$

首先考虑 $M=1$**，定义这个空间的投影方向为D维向量** $\mathbf{u}_1$

出于方便且不失一般性，令 $\mathbf{u}_1^T\mathbf{u}_1=1$

每个数据点 $\mathbf{x}_n$ **在新空间中表示为标量** $\mathbf{u}_1^T\mathbf{x}_n$

样本均值在新空间中表示为 $\mathbf{u}_1^T\bar{\mathbf{x}}$**，其中** $\bar{\mathbf{x}}=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$

投影后样本方差表示为 $\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^T\mathbf{x}_n-\mathbf{u}_1^T\bar{\mathbf{x}}\}^2=\boxed{\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1}$ **最大**

其中原样本方差 $\mathbf{S}=\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\bar{\mathbf{x}})(\mathbf{x}_n-\bar{\mathbf{x}})^T$

● 最大方差思想

使用较少的数据维度保留住较多的原数据特性

目标是**最大化** $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1,\quad s.t.\quad \mathbf{u}_1^T\mathbf{u}_1=1$

利用拉格朗日乘子法 $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1+\lambda_1(1-\mathbf{u}_1^T\mathbf{u}_1)$

对 $\mathbf{u}_1$ 求导置零得到 $\mathbf{S}\mathbf{u}_1=\lambda_1\mathbf{u}_1$

$\mathbf{u}_1$是S的特征向量

进一步得到 $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1=\lambda_1$

$\mathbf{u}_1$是S最大特征值对应的特征向量时
方差取到极大值，称$\mathbf{u}_1$为第一主成分

● 最大方差思想

使用较少的数据维度保留住较多的原数据特性

考虑更一般性的情况($M>1$), 新空间中数据方差最大的最佳投影方向由协方差矩阵S的*M*个特征向量$\mathbf{u}_1,...,\mathbf{u}_M$定义, 其分别对应*M*个最大的特征值 $\lambda_1,...,\lambda_M$

首先获得方差最大的1维，生成该维的补空间；

继续在补空间中获得方差最大的1维，生成新的补空间；

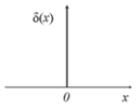
依次循环下去得到*M*维的空间。

● 最小均方误差思想

使原数据与降维后的数据(在原空间中的重建)的误差最小

定义一组正交的*D*维基向量 $\{\mathbf{u}_i\},i=1,...,D$，满足

$$\mathbf{u}_i^T\mathbf{u}_j=\delta_{ij}$$



由于基是完全的，每个数据点可以表示为基向量的线性组合

$$\mathbf{x}_n=\sum_{i=1}^{D}\alpha_{ni}\mathbf{u}_i$$

相当于进行了坐标变换

$$\{\mathbf{x}_{n1},...,\mathbf{x}_{nD}\}\xrightarrow{\{\mathbf{u}_i\}}\{\alpha_{n1},...,\alpha_{nD}\}$$

$$\alpha_{nj}=\mathbf{x}_n^T\mathbf{u}_j$$

● 最小均方误差思想

使原数据与降维后的数据(在原空间中的重建)的误差最小

那么 $\mathbf{x}_n = \sum_{i=1}^{D}(\mathbf{x}_n^T\mathbf{u}_i)\mathbf{u}_i$

在*M*维变量($M < D$)生成的空间中对其进行表示

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i\mathbf{u}_i$$

独特的　　共享的

目标最小化失真度 $J = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$

倒数置零得到 $z_{nj} = \mathbf{x}_n^T\mathbf{u}_j, j = 1, .., M$

$b_j = \bar{\mathbf{x}}^T\mathbf{u}_j, j = M+1, ..., D$

● 最小均方误差思想

使原数据与降维后的数据(在原空间中的重建)的误差最小

有 $\mathbf{x}_n - \tilde{\mathbf{x}}_n = \sum_{i=M+1}^{D}\{(\mathbf{x}_n - \bar{\mathbf{x}})^T\mathbf{u}_i\}\mathbf{u}_i$

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(\mathbf{x}_n^T\mathbf{u}_i - \bar{\mathbf{x}}^T\mathbf{u}_i)^2 = \sum_{i=M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i$$

拉格朗日乘子法 $\tilde{J} = \sum_{i=M+1}^{D}\mathbf{u}_i^T\mathbf{S}\mathbf{u}_i + \sum_{i=M+1}^{D}\lambda_i(1 - \mathbf{u}_i^T\mathbf{u}_i)$

求导得到 $\mathbf{S}\mathbf{u}_i = \lambda_i\mathbf{u}_i$　　***J*最小时取*D-M*个最小的特征值**

对应失真度为 $J = \sum_{i=M+1}^{D}\lambda_i$　　**主子空间对应*M*个最大特征值**

八. 给定方程组形式：$Ya = b$，其中 $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix}$，$y_n$ 是规范化增广向量样本，Y 是 $N \times \hat{d}$ 维矩阵，通常 $N > \hat{d}$，一般为列满秩矩阵，$b = [b_1 \;\; b_2 \;\; \cdots \;\; b_N]$，$b$ 是 N 维向量，$b_n > 1,\; n = 1, 2, \cdots, N$。定义误差向量：$e = Ya - b$ 及平方误差准则函数：$J_S(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{n-1}^{N}\left(a^T y_n - b_n\right)^2$。

问题：请采用解析法求得 a^*，并证明在 $b = \begin{bmatrix} N/N_1 \\ \vdots \\ N/N_1 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix}$ （N_1 个 N/N_1，N_2 个 N/N_2）条件下最小平方误差准则

函数的解 a^* 与 Fisher 线性判别的解相同。（20 分，选做）

- **求使 $J_S(a)$ 最小的 a^* (最小二乘近似解/伪逆解/MSE解)**

采用解析法求伪逆解 $J_S(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{.}^{N}(a^T y_n - b_n)^2$

$$\nabla J_S(a) = \sum_{n-1}^{N} 2(a^T y_n - b_n)y_n = 2Y^T(Ya - b)$$

令 $\nabla J_S(a) = 0$

得 $Y^T Y a^* = Y^T b$ 矩阵 $Y^T Y$ 是 $\hat{d} \times \hat{d}$ 方阵一般非奇异

唯一解 $a^* = (Y^T Y)^{-}1Y^T b = Y^+ b$

其中 $\hat{d} \times N$ **矩阵** $Y^+ = (Y^T Y)^{-}1Y^T$ **是Y的左逆矩阵**

如何选 b?

$$b = \begin{bmatrix} N/N_1 \\ \cdots \\ N/N_1 \\ N/N_2 \\ \cdots \\ N/N_2 \end{bmatrix} \begin{matrix} N_1\text{个} \\ \\ \\ N_2\text{个} \end{matrix}$$

a^* 等价于Fisher解

$$g_0(x) = P(w_1|x) - P(w_2|x)$$

$N \to \infty, b = [1, 1, \cdots, 1]^T$（$N$个） 以最小均方误差逼近贝叶斯判别函数

九. 通过含有一层隐藏层的神经网络推导 BP 反传算法。（20 分，选做）

其中：

给定训练集 $D = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$，$\mathbf{x}_i \in \mathbb{R}^D, \mathbf{y}_i \in \mathbb{R}^K$ 。

激活函数为 Sigmoid 函数：$\text{sigmoid}(x) = \dfrac{1}{1+e^{-x}}$ 。

准则函数为 $E(\mathbf{w}) = \dfrac{1}{2}\sum_{n=1}^{N}\{\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{y}_n\}^2$，$\mathbf{y}(\mathbf{x}_n, \mathbf{w})$ 表示第 n 个样本 $\mathbf{x}_n$ 对应的神经网络输出向量。

前馈网络结构包含 D 个输入神经元，K 个输出神经元，以及 M 个隐层神经元，如图 2 所示。

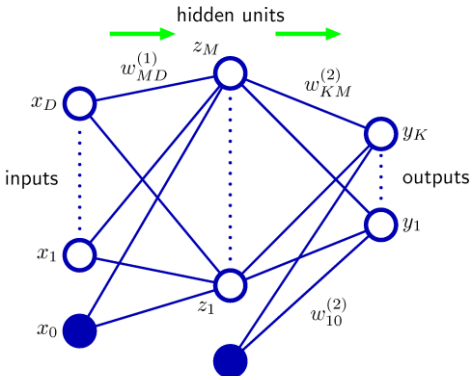
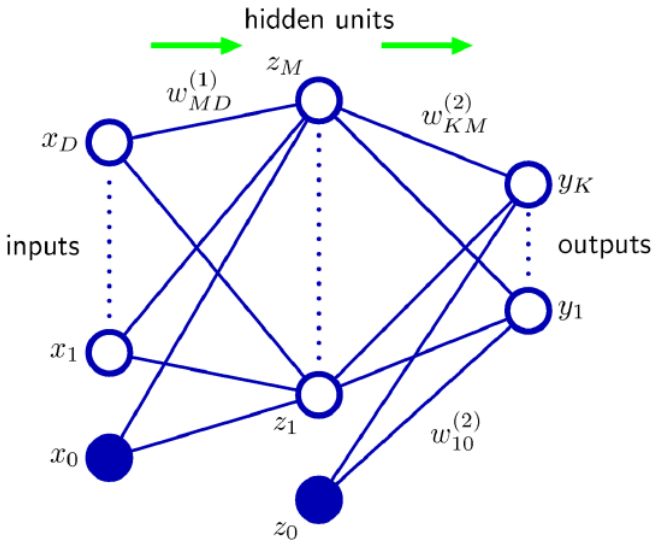


图 2 前馈网络结构图



$$a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i$$

$$z_j = h(a_j)$$

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j$$

$$y_k = \sigma(a_k)$$

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{i=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$

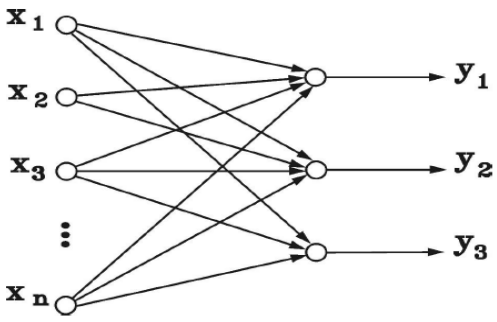
● 定义准则函数

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2$$

寻找一个 $\mathbf{w}$，使得 $E(\mathbf{w})$ 最小。

$$\nabla E(\mathbf{w}) = 0$$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)})$$



$$y_k = \sum_i w_{ki} x_i$$

$$E_n = \frac{1}{2}\sum_k (y_{nk} - t_{nk})^2$$

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w})$$

误差

$$\frac{\partial E_n}{\partial w_{jk}} = (y_{nk} - t_{nk}) x_{nk}$$

$$E_n = \frac{1}{2}\sum_k (y_{nk} - t_{nk})^2 \qquad a_j = \sum_i w_{ji} z_i \qquad z_j = h(a_j)$$

$$\delta_j \equiv \frac{\partial E_n}{\partial a_j} \qquad \frac{\partial a_j}{\partial w_{ji}} = z_i \qquad \frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

对于输出层而言：$\delta_k = y_k - t_k$

对于隐藏层而言：$\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j}$

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$$

- 初始化权重w_{ij}
- 对于输入的训练样本，求取每个节点输出和最终输出层的输出值
- 对输出层求取 $\delta_k = y_k - t_k$
- 对于隐藏层求取 $\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$
- 求取输出误差对于每个权重的梯度 $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$
- 更新权重 $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{\tau} + \eta\Delta E(\mathbf{w}^{(\tau)})$

多层前馈网络的学习算法比较复杂，其主要困难是中间的隐层不直接与外界连接，无法直接计算其误差。反向逐层传播输出层的误差，以间接算出隐层误差。算法分为两个阶段：第一阶段（正向过程）输入信息从输入层经隐层逐层计算各单

元的输出值；第二阶段(反向传播过程)由输出误差逐层向前算出隐层各单元的误差，并用此误差修正前层权值。