# 模式识别 第二十二周读书报告 第五部分 高阶课题

## 第 13 章 正态分布

### 13.1 定义

**单变量正态分布**

单变量的正态分布的概率密度函数：

$$\mathrm{p(x)}=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}$$

其中μ是期望值，$\sigma^2$是方差。符号$\mathrm{X}\sim\mathrm{N}(\mu,\sigma^2)$。

标准正态就是期望（均值）0 和方差 1。符号$\mathrm{X}\sim\mathrm{N}(0,1)$。

**多元正态分布**

多元正态分布的概率密度函数:

$$p(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

其中 x 是一个 d 维向量，$\mu$是 d 维均值，Σ是一个 d*d 的协方差矩阵。

### 13.2 线性运算与求和

**单变量情形**

假设$X_1\sim N(\mu_1,\sigma_1^2)$和$X_2\sim N(\mu_2,\sigma_2^2)$是两个独立的单变量正态分布变量，显然有

$$aX_1+b\sim N(a\mu_1+b,a^2\sigma_1^2),$$

其中 a 和 b 是两个常数。

现考虑一个随机变量$Z=X_1+X_2$，Z 的概率密度可以通过一个卷积来计算，即

$$p_Z(z)=\int_{-\infty}^{+\infty}p_{x_1}(x_1)p_{x_2}(z-x_1)dx_1$$

并且$\mathrm{Z}\sim\mathrm{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$。

**多变量的情形**

假设$X\sim N(\mu,\Sigma)$是一个 d 维正态随机变量，A 是一个有 q 行、d 列的矩阵,b 是一个 q 维的向量，那么$Z=AX+b$是一个 q 维正态随机变量

$$Z\sim N(A\mu+b,A\Sigma A^T)$$

Z 的特征方程为

$$\varphi_Z(t)=exp(it^T(A\mu+b)-\frac{1}{2}t^T(A\Sigma A^T)t)$$

假设$X_1\sim N(\mu_1,\sum_1)$和$X_2\sim N(\mu_2,\sum_2)$是两个独立的 d 维正态随机变量，有$\mathrm{Z}=X_1+X_2$

$$\mathrm{Z}\sim\mathrm{N}(\mu_1+\mu_2,\sum_1+\sum_2)$$

此时 Z 的特征方程为

$$\varphi_Z(t)=exp(it^T(\mu_1+\mu_2)-\tfrac{1}{2}t^T(\Sigma_1+\Sigma_2)t)$$

### 13.3 几何和马氏距离

下图展示了一个二元正态概率密度函数，正态分布概率密度函数只有一个峰，对应于均值向量，密度函数的形状有协方差矩阵决定

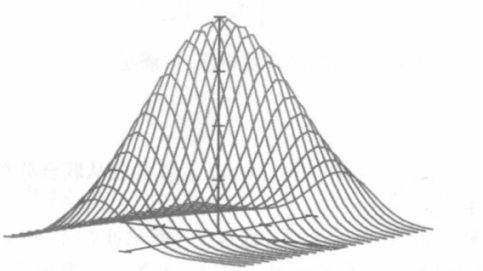


图 13.1　二元正态概率密度函数 (见彩插)

下图展示了二元正态随机变量的概率等高线，在同一条概率等高线上的所有点必须在下式中计算得到一个常数:

$$r^2(x,\mu) = (x-\mu)^T\Sigma^{-1}(x-\mu) = c$$

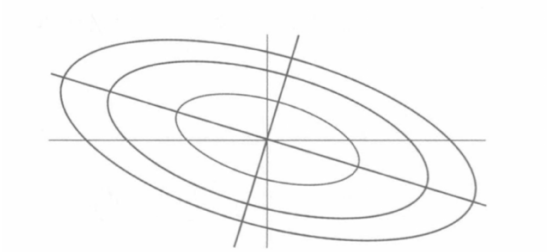


图 13.2　一个二元正态分布的概率等高线

**13.4 条件作用**

假设$X_1$和$X_2$是两个多元正态随机变量，其联合概率密度函数是

$$p(\begin{bmatrix}x_1\\x_2\end{bmatrix}) = \frac{1}{(2\pi)^{(d_1+d_2)/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}\begin{bmatrix}x_1-\mu_1\\x_2-\mu_2\end{bmatrix}^T\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix}x_1-\mu_1\\x_2-\mu_2\end{bmatrix})$$

其中$d_1$和$d_2$分别是$X_1$和$X_2$的维度，并且

$$\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

矩阵$\Sigma_{12}$和$\Sigma_{21}$是$\mathrm{x}_1$和$\mathrm{x}_2$之间的协方差矩阵，并满足

$$\Sigma_{12} = (\Sigma_{21})^T$$

$\sum_{11}$的舒尔补，定义为

$$\mathrm{S}_{11} = \sum_{22} - \sum_{21}\sum_{11}^{-1}\sum_{12}$$

则有

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1} = \begin{bmatrix}S_{22}^{-1} & -S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma_{12}^T\Sigma_{22}^{-1} & \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{12}^TS_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{bmatrix}$$

$$\begin{aligned}&\begin{bmatrix}x_1-\mu_1\\x_2-\mu_2\end{bmatrix}^T\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix}x_1-\mu_1\\x_2-\mu_2\end{bmatrix}\\ &= x_1'S_{22}^{-1}x_1' + x_2'^T(\Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{12}^TS_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1})x_2'\\ &= (x_1'+\Sigma_{12}\Sigma_{22}^{-1}x_2')^TS_{22}^{-1}(x_1'+\Sigma_{12}\Sigma_{22}^{-1}x_2') + x_2'^T\Sigma_{22}^{-1}x_2'.\end{aligned}$$

因此联合分布可以写为

$$p(\begin{bmatrix}x_1\\x_2\end{bmatrix}) = \frac{1}{(2\pi)^{(d_1+d_2)/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}x_2'^T\Sigma_{22}^{-1}x_2')$$

条件概率$\mathrm{p}(\mathrm{x}_1|\mathrm{x}_2)$如下

$$\mathrm{p}(\mathrm{x}_1|\mathrm{x}_2)=\frac{1}{(2\pi)^{\mathrm{d}_1}|\mathrm{S}_{22}|^{1/2}}\exp\left(-\frac{(\mathrm{x}_1'-\sum_{12}\sum_{22}^{-1}\mathrm{x}_2')^{\mathrm{T}}\mathrm{S}_{22}^{-1}(\mathrm{x}_1'-\sum_{12}\sum_{22}^{-1}\mathrm{x}_2')^{\mathrm{T}}}{2}\right)$$

或

$$\begin{aligned}\mathrm{x}_1|\mathrm{x}_2 &\sim \mathrm{N}(\mu_1+\sum_{12}\sum_{22}^{-1}\mathrm{x}_2',\mathrm{S}_{22}^{-1})\\ &\sim \mathrm{N}(\mu_1+\sum_{12}\sum_{22}^{-1}(\mathrm{x}_2-\mu_2),\sum_{11}-\sum_{12}\sum_{22}^{-1}\sum_{21})\end{aligned}$$

**13.5 高斯分布的乘积**

假设$X_1$和是两个独立的 d 维正态随机变量，与这两个正态分布密度的乘积成正比的概率密度函数如下

$$p_x(x)=\alpha p_1(x)p_2(x)$$

其中$\alpha$是一个合适的规范化常数,使得$p_x(x)$是一个有效的密度函数

将两个正态密度写成正则形式:

$$p_1(x)=exp(\alpha_1+\eta_1^Tx-\frac{1}{2}x^T\Lambda_1x)$$

$$p_2(x)=exp(\alpha_2+\eta_2^Tx-\frac{1}{2}x^T\Lambda_2x)$$

那么$p_x(x)$的密度容易计算

$$p_X(x)=\alpha p_1(x)p_2(x)$$

$$=exp(\alpha'+(\eta_1+\eta_2)^Tx-\frac{1}{2}x^T(\Lambda_1+\Lambda_2)x)$$

扩展到 n 个正态分布，那么$p_X(x)=\alpha\prod_{i=1}^n p_i(x)$也是一个正态密度，由下式给出

$$p_X(x)=exp\left(\alpha'+\left(\sum_{i=1}^n\eta_i\right)^Tx-\frac{1}{2}x^T\left(\sum_{i=1}^n\Lambda_i\right)x\right)$$

使用矩参数化形式

$$\mathrm{p}(\mathrm{x})=\mathrm{N}(\mathrm{x};\mu,\sum)$$

其中

$$\sum^{-1}=\sum_1^{-1}+\sum_2^{-1}+\cdots+\sum_n^{-1}$$

$$\sum^{-1}\mu=\sum_1^{-1}\mu_1+\sum_2^{-1}\mu_2+\cdots+\sum_n^{-1}\mu_n$$

**13.6 应用：卡尔曼滤波**

**模型：**

给定动态模型：

$$x_k=Ax_{k-1}+w_{k-1}$$

给定线性测量模型：

$$z_k=Hx_k+v_k$$

假设过程噪声$w_k$和测量噪声$v_k$是正态的且和其他所有随机变量独立

$$\mathrm{w}\sim\mathrm{N}(0,\mathrm{Q})$$

$$\mathrm{v}\sim\mathrm{N}(0,\mathrm{R})$$

**估计：**

卡尔曼滤波可以被分为两个步骤：

1、根据估计$\mathrm{p}(\mathrm{x}_{\mathrm{k}-1})$和动态模型，我们可以得到一个估计$\mathrm{p}(x_k^-)$。负号代表该估计是将观测值纳入考虑之前完成的。

2、根据$\mathrm{p}(x_k^-)$和测量模型，我们得到最后的估计$\mathrm{p}(\mathrm{x}_\mathrm{k})$。这个估计事实上是以观测值$z_k$

和之前的状态$x_{t-1}$为条件的。

假设在 k-1 时刻，我们已经得到的估计是正态分布

$$p(x_{k-1})\sim N(\mu_{k-1}, P_{k-1})$$

在动态模型中应用线性运算的公式，我们可以得到$x_k^-$的估计

$$x_k^- \sim N(\mu_k^-, P_k^-)$$

$$\mu_k^- = A\mu_{k-1}$$

$$P_k^- = AP_{k-1}A^T + Q$$

将观测值$z_k$为条件作用到估计$p(x_k^-)$上给出了我们想要得到的估计$p(x_k)$，不考虑 k 时刻的观测值，最好的估计是

$$Hx_k^- + v_k$$

$z_k, x_k^-$的联合协方差矩阵为

$$\begin{bmatrix} P_k^- & P_k^- H^T \\ HP_k^- & HP_k^- H^T + R \end{bmatrix}$$

应用条件作用的性质，可得

$$p(x_k) = p(x_k^- | z_k) \sim N(\mu_k, P_k))$$

$$P_k = P_k^- - P_k^- H^T (HP_k^- H^T + R)^{-1} HP_k^-$$

$$\mu_k = \mu_k^- + P_k^- H^T (HP_k^- H^T + R)^{-1}(z_k - H\mu_k)$$

定义$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$，上述公式被简化为

$$P_k = (I - K_k H)P_k^-$$

$$\mu_k = \mu_k^- + K_k(z_k - H\mu_k)$$

项$K_k$被称为卡尔曼增益矩阵，项$z_k - H\mu_k$被称为新息。

# 第 14 章 EM 算法的基本思想

## 14.1 GMM 高斯混合模型

若$p_1$、$p_2$和$p_3$是三个不同的概率密度函数，且都服从高斯分布。以下图为例，$p_3(x) = 0.2p_2(x) + 0.8p_3(x)$。由于权重非负且和为 1，$p_3$是两个高斯（$p_1$和$p_2$）的混合，因此是一个高斯混合模型。

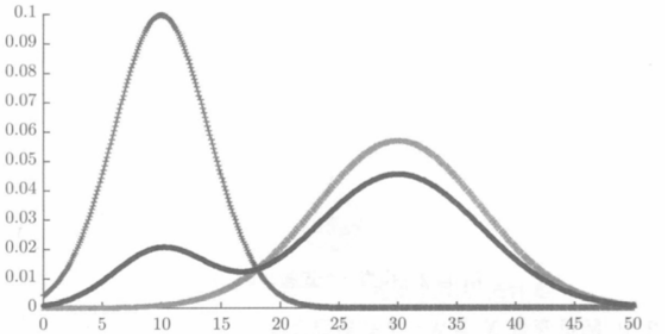



图 14.1 一个简单的 GMM 图示 (见彩插)

一个 GMM 是一个分布，其概率密度函数为

$$p(x) = \sum_{i=1}^{N} \alpha_i N(x; \mu_i, \Sigma_i) = \sum_{i=1}^{N} \frac{\alpha_i}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} exp(-\frac{1}{2}(x-\mu_i)^T {\Sigma_i}^{-1}(x-\mu_i))$$

其中 x 是一个 d 维随机向量，第 i 个分量的权重是$\alpha_i$（称为混合系数），第 i 个分量有均值向量$\mu_i$和协方差矩阵$\sum_i$。

混合系数必须满足如下条件：

$$\sum_{i=1}^{N} \alpha_i = 1$$

$$\alpha_i \geq 0，\ \forall i$$

**14.2 基于隐变量的诠释**

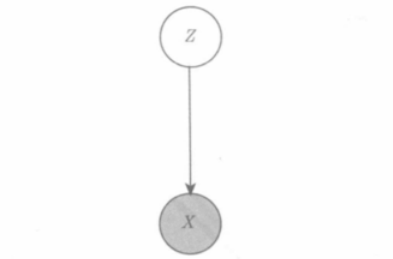


图 14.2 作为一个图模型的 GMM

随机变量 X 服从一个高斯混合模型，它的参数是

$$\theta = \{\alpha_i, \mu_i, \sum_i, \}_{i=1}^{N}$$

我们可以从上式的概率密度函数中采样一个 GMM 实例。

还有另一种方法可以用来进行采样。

我们定义一个随机变量 Z，Z 是一个离散多项分布，从$\{1,2,\cdots,N\}$中取值。Z 取值为$Z = i$的概率是$\alpha_i$。两步采样的步骤如下

1、从 Z 中采样，得到值 i（$1 \leq i \leq N$）；

2、从第 i 个高斯分量$N(\mu_i, \sum_i)$中采样得到$x$。

容易验证，从这两步中得到的 x 服从上面公式的 GMM 分布。

**14.3 EM 算法的非正式描述**

- 首先以任何合理的方式初始化 θ 的值;
- 然后，使用 X 和 θ 的当前估计值，可以估计最可能的 Z(或其后验分布的期望);
- 基于这个 Z 的估计,可以使用 X 找到 θ 的一个更好的估计;
- 一个更好的 θ(结合 X)可以导致 Z 的更好估计;
- 这个过程(交替估计 θ 和 Z)可以一直进行,直到 θ 的变化很小 (即,当该过程收敛). 使用更加非正式的语言，在适当的参数初始化后,可以执行:

**E 步（期望步骤）** 使用数据和当前参数值,找到一个对不可观测的隐变量更好的猜测;

**M 步（最大化步骤）** 使用当前对隐变量的猜测和数据,找到一个更好的参数估计;

**重复** 重复以上两步骤直至收敛

**14.4 期望最大化算法**

可观测变量 X 和隐变量 Z 的联合概率密度是$p(X, Z; \theta)$,其中 θ 是参数，现有一组 X 的实例

$$X = \{x_1, x_2, \ldots, x_M\}$$

以用来学习参数，任务是从 X 中估计 θ

对每一个$x_j$,有一个对应的$z_j$，记作

$$Z = \{z_1, z_2, \ldots, z_M\}$$

现在θ包括了对应于 Z 的参数，在 GMM 的例子中，$\{z_i\}_{i=1}^{M}$是对 Z 的估计，$\{\alpha_i, \mu_i, \sum_i\}_{i=1}^{N}$是指定 X 的参数，θ包含所有这两组参数。

### 14.4.1 联合非凹的不完整数据对数似然

如果一个最小化问题是凸的，一般可以认为这个问题容易，但是非凸问题很难求解。凹最大化问题通常被认为是容易的，非凹最大化问题通常是困难的。

$p(\chi|\theta)$的似然是

$$p(\chi|\theta)=\prod_{j=1}^{M}(\sum_{i=1}^{N}\frac{\alpha_i}{(2\pi)^{d/2}|\sum_i|^{1/2}}\exp(-\frac{1}{2}(x_j-\mu_i)^T\sum_i^{-1}(x_j-\mu_i)))$$

对数似然有如下形式

$$\sum_{j=1}^{M}ln(\sum_{i=1}^{N}\frac{\alpha_i}{(2\pi)^{d/2}|\sum_i|^{1/2}}\exp(-\frac{1}{2}(x_j-\mu_i)^T\sum_i^{-1}(x_j-\mu_i)))$$

这个公式对联合的优化变量$\{\alpha_i,\mu_i,\sum_i\}_{i=1}^{n}$非凹，换而言之，这是困难的最大化问题。

由于我们有两组随机变量 X 和 Z，但是 Z 不在公式中，因此上式中的对数似然被称为不完整数据对数似然。

### 14.4.2 (可能是)凹的完整数据对数似然

完整数据对数似然是

$$lnp(\mathcal{X},\mathcal{Z}|\theta)$$

其中假设隐变量 Z 是已知的，知道隐变量通常会优化问题。例如完整数据对数似然有可能变成了凹的。

在 GMM 中，$z_i$向量（构成了 Z）是一个 N 维向量，其中有 N-1 个 0，仅有一维为 1。因此完整数据似然是

$$p(\mathcal{X},\mathcal{Z}|\theta)=\prod_{j=1}^{M}\prod_{i=1}^{N}\left[\frac{\alpha_i}{(2\pi)^{d/2}|\sum_i|^{1/2}}\exp(-\frac{1}{2}(x_j-\mu_i)^T\sum_i^{-1}(x_j-\mu_i))\right]^{z_{ij}}$$

这个公式可以用两步采样的过程进行解释。

完整数据对数似然是

$$\sum_{j=1}^{M}\sum_{i=1}^{N}z_{ij}\left(\frac{1}{2}(ln|\sum_i^{-1}|-(x_j-\mu_i)^T\sum_i^{-1}(x_j-\mu_i))+ln\alpha_i\right)+const$$

其中$const$指不受参数θ影响的各项常数。

### 14.4.3 EM 算法

1、 令 t = 0；

2、初始化参数$\theta^{(0)}$；

3、E （期望）步：计算$p(\mathcal{Z}|\mathcal{X},\theta^{(t)})$；

4、M（最大化）步：1）找到期望

$$Q(\theta,\theta^{(t)})=\mathbb{E}_{\mathcal{Z}|\mathcal{X},\theta^{(t)}}[lnp(\mathcal{X},\mathcal{Z}|\theta)]$$

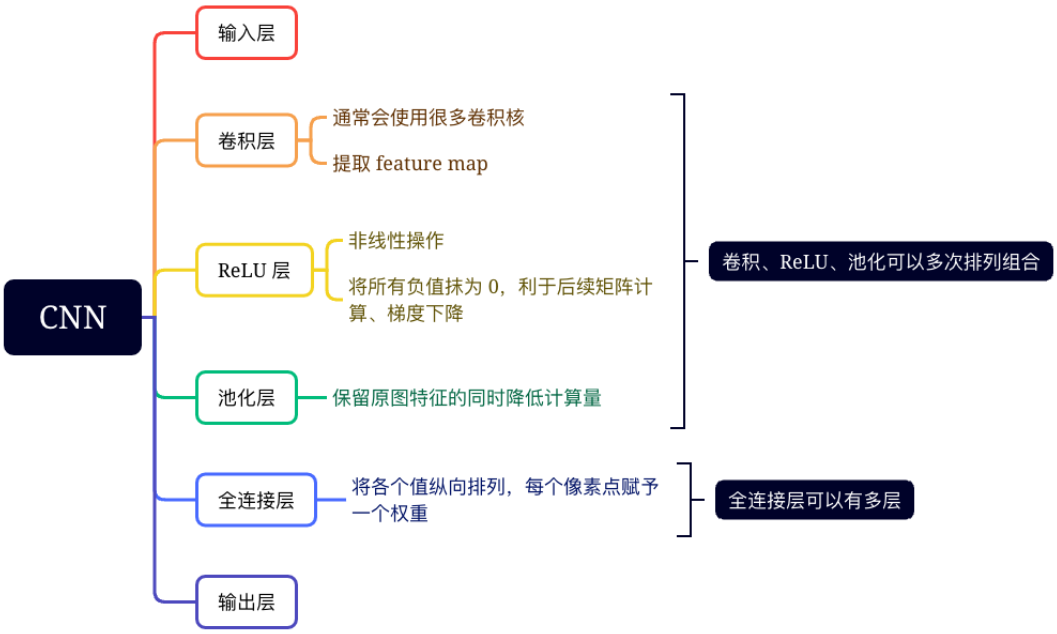
5、M（最大化）步：2）找到一个新的参数估计

$$\theta^{(t+1)}=\arg\max_{\theta}Q\left(\theta,\theta^{(t)}\right)$$

6、令 t = t + 1 ；

7、若对数似然没有收敛，再次前往 E 步执行。

# 第 15 章 卷积神经网络




## 15.1 随机梯度下降

我们用到一种名为梯度下降（gradient descent）的方法， 这种方法几乎可以优化所有深度学习模型。 它通过不断地在损失函数递减的方向上更新参数来降低误差。

梯度下降最简单的用法是计算损失函数（数据集中所有样本的损失均值） 关于模型参数的导数（在这里也可以称为梯度）。 但实际中的执行可能会非常慢：因为在每一次更新参数之前，我们必须遍历整个数据集。 因此，我们通常会在每次需要计算更新的时候随机抽取一小批样本， 这种变体叫做小批量随机梯度下降（minibatch stochastic gradient descent）。

我们用下面的数学公式来表示这一更新过程

$$(\mathbf{w}, b) \leftarrow (\mathbf{w}, b) - \frac{\eta}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \partial_{(\mathbf{w},b)} l^{(i)}(\mathbf{w}, b)$$

算法的步骤如下： （1）初始化模型参数的值，如随机初始化； （2）从数据集中随机抽取小批量样本且在负梯度的方向上更新参数，并不断迭代这一步骤。

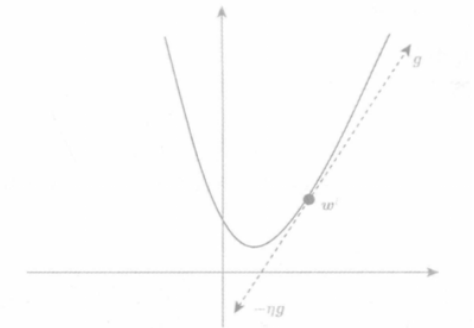



图 15.1 梯度下降方法图示, 其中 $\eta$ 是学习率 (见彩插)

## 15.2 反向传播算法

神经网络会得出一个结果，随后会将这个结果与真实结果进行比较，若预测结果错误，则会根据损失函数，通过修改卷积核的参数或每个神经元的权重，来使得损失函数最小。而误差又是一层一层反馈上去的，因此称为反向传播。

### 15.3 参数共享机制

在卷积层中的每个神经元连接数据窗的权重是固定的，每个神经元只关注一个特性。神经元就是图像处理中的滤波器，比如边缘检测中的 Sobel 滤波器。卷积层中的每个滤波器都会有自己所关注的一个图像特征，比如垂直边缘，水平边缘，颜色和纹理等等，这些所有的神经元加起来就好比是整张图像的特征提取器的集合。

### 15.4 克罗内克积

给定两个矩阵 A、B，克罗内克积定义为分块矩阵：

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

克罗内克积的性质，对有适当大小的矩阵 A、X、B，有

$$(A \otimes B)^T = A^T \otimes B^T,$$
$$\operatorname{vec}(AXB) = (B^T \otimes A)\operatorname{vec}(X),$$

vec()运算是将张量展开为一个向量。

### 15.5 VGG-16 网络

VGG-16 模型的结构如下表

表 15.2 VGG-Verydeep-16的结构与感受野

| | 类型 | 描述 | 感受野 | | 类型 | 描述 | 感受野 |
|---|---|---|---|---|---|---|---|
| 1 | Conv | 64;3x3;p=1,st=1 | 212 | 20 | Conv | 512;3x3;p=1,st=1 | 20 |
| 2 | ReLU | | 210 | 21 | ReLU | | 18 |
| 3 | Conv | 64;3x3;p=1,st=1 | 210 | 22 | Conv | 512;3x3;p=1,st=1 | 18 |
| 4 | ReLU | | 208 | 23 | ReLU | | 16 |
| 5 | Pool | 2x2;st=2 | 208 | 24 | Pool | 2x2;st=2 | 16 |
| 6 | Conv | 128;3x3;p=1,st=1 | 104 | 25 | Conv | 512;3x3;p=1,st=1 | 8 |
| 7 | ReLU | | 102 | 26 | ReLU | | 6 |
| 8 | Conv | 128;3x3;p=1,st=1 | 102 | 27 | Conv | 512;3x3;p=1,st=1 | 6 |
| 9 | ReLU | | 100 | 28 | ReLU | | 4 |
| 10 | Pool | 2x2;st=2 | 100 | 29 | Conv | 512;3x3;p=1,st=1 | 4 |
| 11 | Conv | 256;3x3;p=1,st=1 | 50 | 30 | ReLU | | 2 |
| 12 | ReLU | | 48 | 31 | Pool | | 2 |
| 13 | Conv | 256;3x3;p=1,st=1 | 48 | 32 | FC | (7x7x512)x4096 | 1 |
| 14 | ReLU | | 46 | 33 | ReLU | | |
| 15 | Conv | 256;3x3;p=1,st=1 | 46 | 34 | Drop | 0.5 | |
| 16 | ReLU | | 44 | 35 | FC | 4096x4096 | |
| 17 | Pool | 2x2;st=2 | 44 | 36 | ReLU | | |
| 18 | Conv | 512;3x3;p=1,st=1 | 22 | 37 | Drop | 0.5 | |
| 19 | ReLU | | 20 | 38 | FC | 4096x1000 | |
| | | | | 39 | $\sigma$ | (softmax 层) | |

上述模型有六种类型的层：卷积层、ReLU 层、池化层、全连接层、Dropout 层、Softmax 层。

**感受野**

感受野指的是一个特定的 CNN 特征（特征图上的某个点）在输入空间所受影响的区域。

第一层卷积层的输出特征图像素的感受野的大小等于滤波器的大小；

深层卷积层的感受野大小和它之前所有层的滤波器大小和步长有关系；

计算感受野大小时，忽略了图像边缘的影响，即不考虑 padding 的大小。

感受野的计算是逐层进行的，从输入层开始，逐层迭代到最后输出：

$$l_k = l_{k-1} + [(f_k - 1) * \prod_{i=1}^{k-1} s_i]$$

其中$l_{k-1}$为第 k-1 层的感受野的大小，$f_k$为第 k 层卷积核的大小，或是池化层池化尺寸的大小。