

数据预处理、特征工程、模型评价

李波

研究数据的相关方向包括：

- 机器学习 (machine learning)
- 数据科学 (data science)
- 数据分析 (data analysis)
- 统计学习 (statistical learning)
- 数据挖掘 (knowledge mining)
- 模式识别 (pattern recognition)

“数据” 这个概念包含：

- 文本
- 数学数字
- 图数据
- 表格
- 交易流水
- 图片
- 视频
- 音频

.....

一些数据的例子：

- 谷歌公司每天处理24 PB数据
- Facebook网站每天上传1千万张照片
- Youtube每秒钟上传1小时时长视频
- Twitter每天新增4亿新twitter
- 全世界的卫星每天产生数据都是几百个PB。

截止2020年，全球数据达到44ZB （44000亿GB）

数据预处理

数据预处理将给定数据转换为机器学习可以处理的数据。

- **删除无信息量的特征:** 与任务无关特征应该删除。
- **平衡数据:** 保证每一类别数据个数差不多。
- **补全缺失数据。**
- **删除野值:** 野值是指与同类别其他数据分布规律不一致的数据。
- **非数值数据转换为数值数据:** 机器学习模型只能处理数值型数据。

否 \rightarrow 0, 是 \rightarrow 1

- **数据归一化:** 保证每个特征数值的大小都差不多

$$x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_{.j}}{s_j} \text{ or } x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_{.j}}{\max_j |x_{ij} - \bar{x}_{.j}|}$$

其中 $\bar{x}_{.j} = \frac{\sum_i x_{ij}}{n_0 + n_1}$, $s_j^2 = \frac{\sum_i (x_{ij} - \bar{x}_{.j})^2}{n_0 + n_1 - 1}$ 。

数据预处理

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12842113	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y

数据预处理

非数值特征

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12842113	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y

无用特征

格式错误

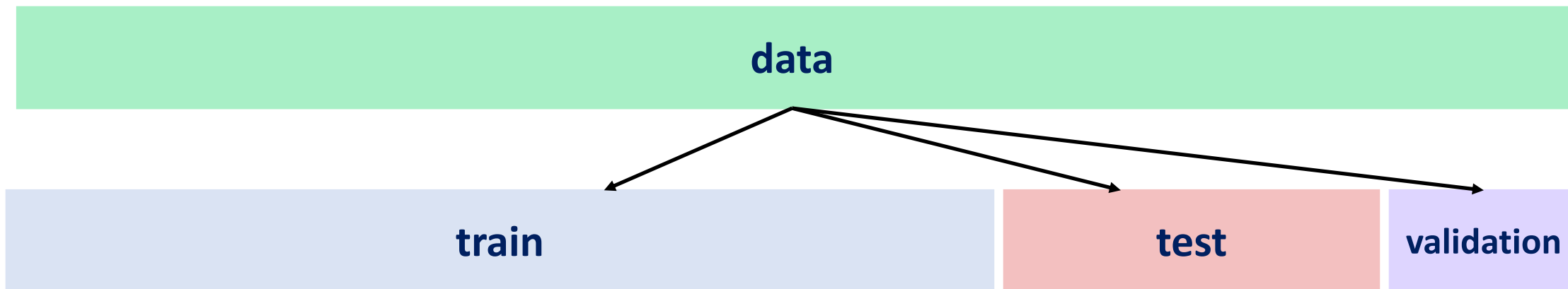
数值缺失

野值

数据不平衡

数据划分

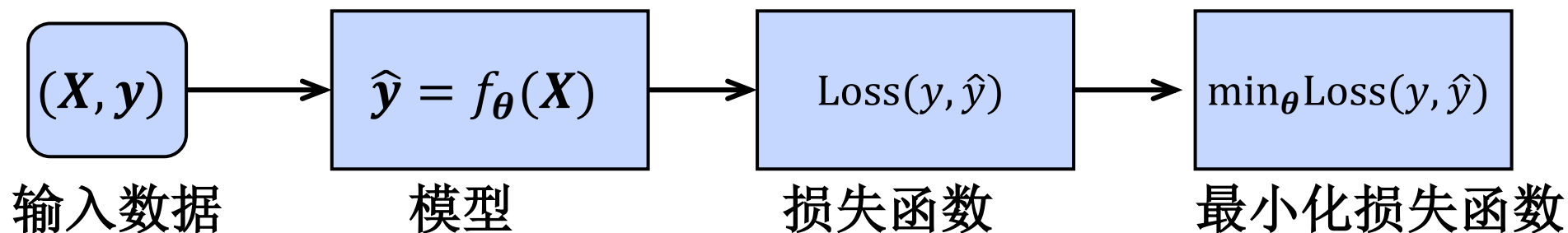
- **训练数据**(training data)
 - 70%，用于训练模型，调整模型参数。
- **测试数据**(testing data)
 - 20%，用于测试模型性能。
- **验证数据**(validation data)
 - 10%，用于确定模型的超参数 (hyper-parameters)



机器学习流程

以有监督学习为例

训练过程



机器学习三要素：

- **模型或者机器学习算法**：选择一个适合问题的模型。
- **损失函数**：如果 $y = \hat{y}$ ，损失函数 $\text{Loss}(y, \hat{y})$ 最小。
- **优化**：优化算法用于调整模型参数，最小化损失函数。

机器学习流程

有监督学习



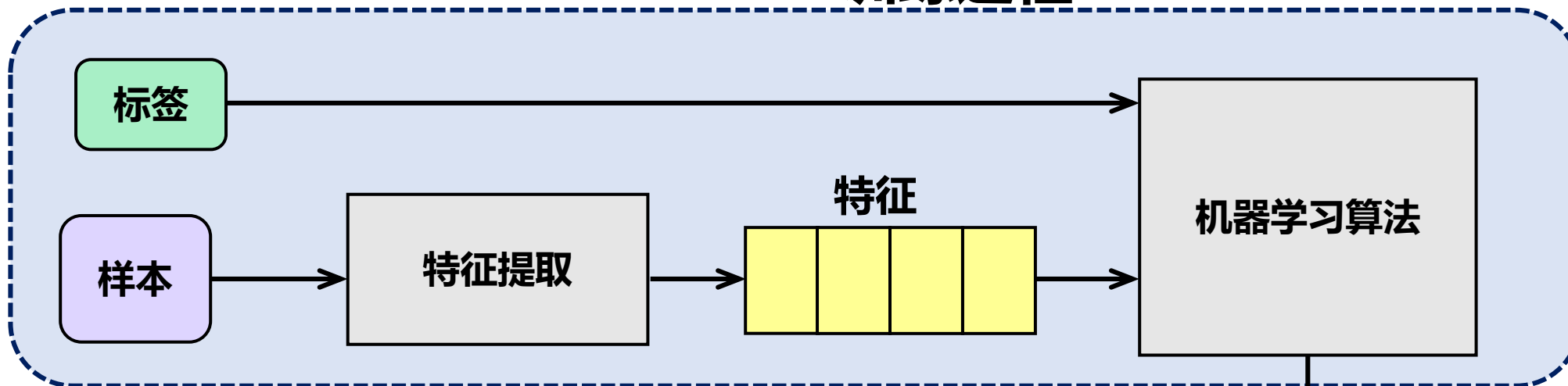
机器学习三要素：

- **模型或者机器学习算法：** 选择一个适合问题的模型。
- **损失函数：** 如果 $y = \hat{y}$ ，损失函数 $\text{Loss}(y, \hat{y})$ 最小。
- **优化：** 优化算法用于调整模型参数，最小化损失函数。

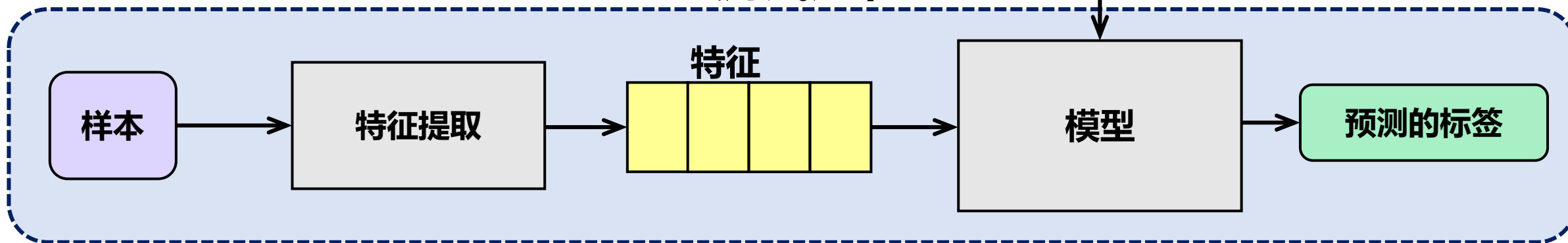
机器学习流程

以有监督学习为例

训练过程



测试过程



特征工程

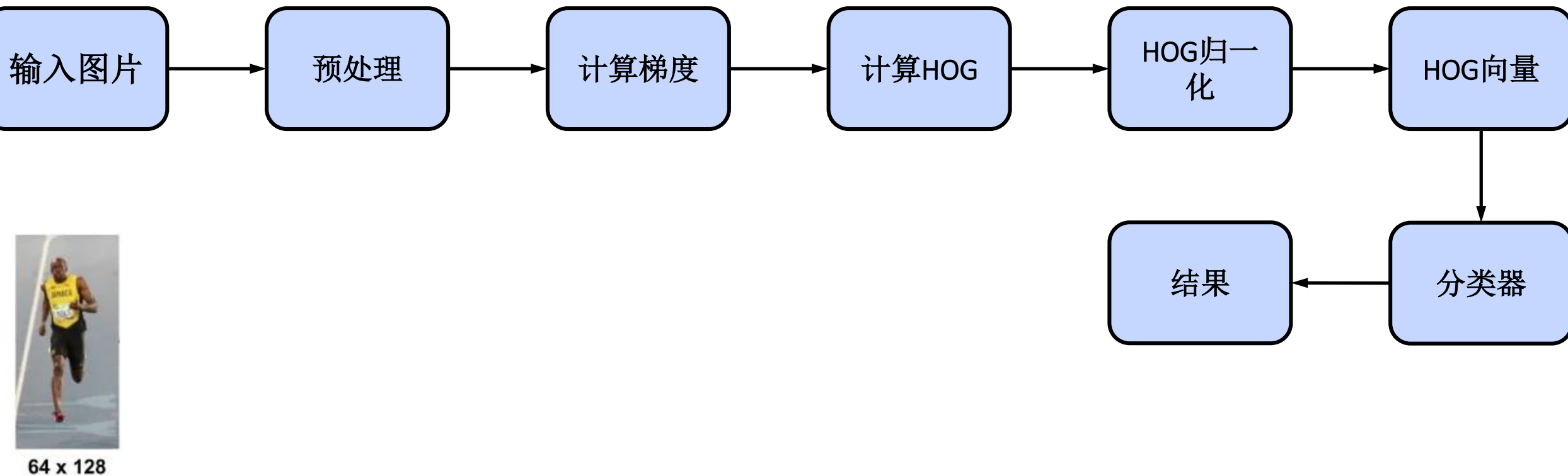


人看到的图像

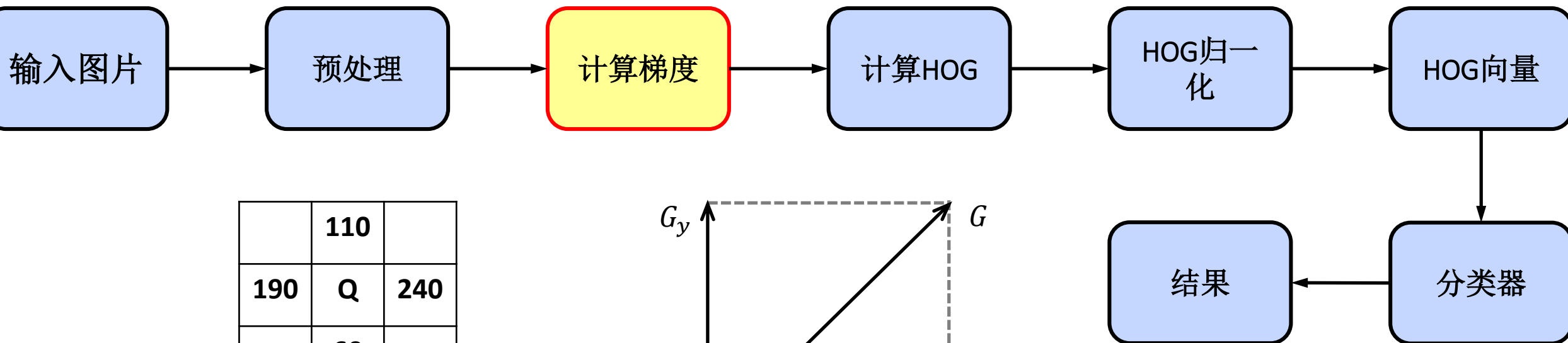
[(226, 137, 125), (226, 137, 125), (223, 137, 133), (223, 136, 128), (226, 138, 120), (226, 129, 116), (228, 138, 123), (227, 134, 124), (227, 140, 127), (225, 136, 119), (228, 135, 126), (225, 134, 121), (223, 130, 108), (226, 139, 119), (223, 135, 120), (221, 129, 114), (221, 134, 108), (221, 131, 113), (222, 138, 121), (222, 139, 114), (223, 127, 109), (223, 132, 105), (224, 129, 102), (221, 134, 109), (218, 131, 110), (221, 133, 113), (223, 130, 108), (225, 125, 98), (221, 130, 121), (221, 129, 111), (220, 127, 121), (223, 131, 109), (225, 127, 103), (223, 134, 109), (226, 128, 106), (223, 135, 122), (225, 133, 112), (227, 144, 124), (229, 135, 104), (231, 142, 123), (231, 143, 116), (232, 142, 112), (230, 143, 117), (233, 150, 121), (234, 148, 121), (237, 154, 123), (233, 153, 121), (231, 149, 121), (237, 149, 119), (238, 149, 116), (234, 143, 118), (235, 154, 122), (234, 145, 116), (232, 142, 121), (233, 135, 112), (230, 133, 121), (227, 118, 98), (221, 120, 105), (219, 127, 127), (213, 110, 109), (203, 98, 103), (202, 82, 91), (187, 86, 98), (174, 74, 92), (169, 63, 84), (166, 65, 85), (158, 63, 89), (153, 61, 96), (155, 55, 80), (157, 67, 98), (166, 72, 93), (161, 66, 82), (166, 73, 89), (165, 69, 87), (169, 73, 91), ...]

计算机看到的图像

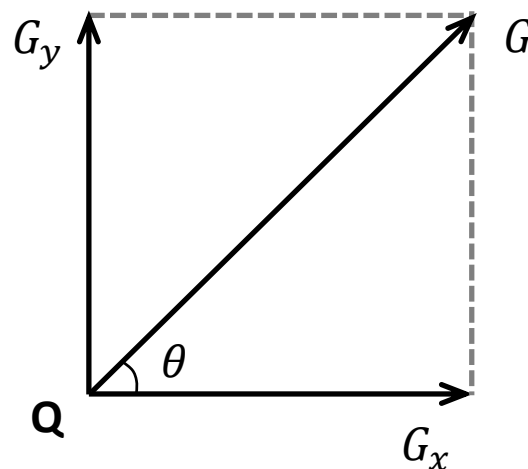
特征工程



特征工程



	110	
190	Q	240
	60	

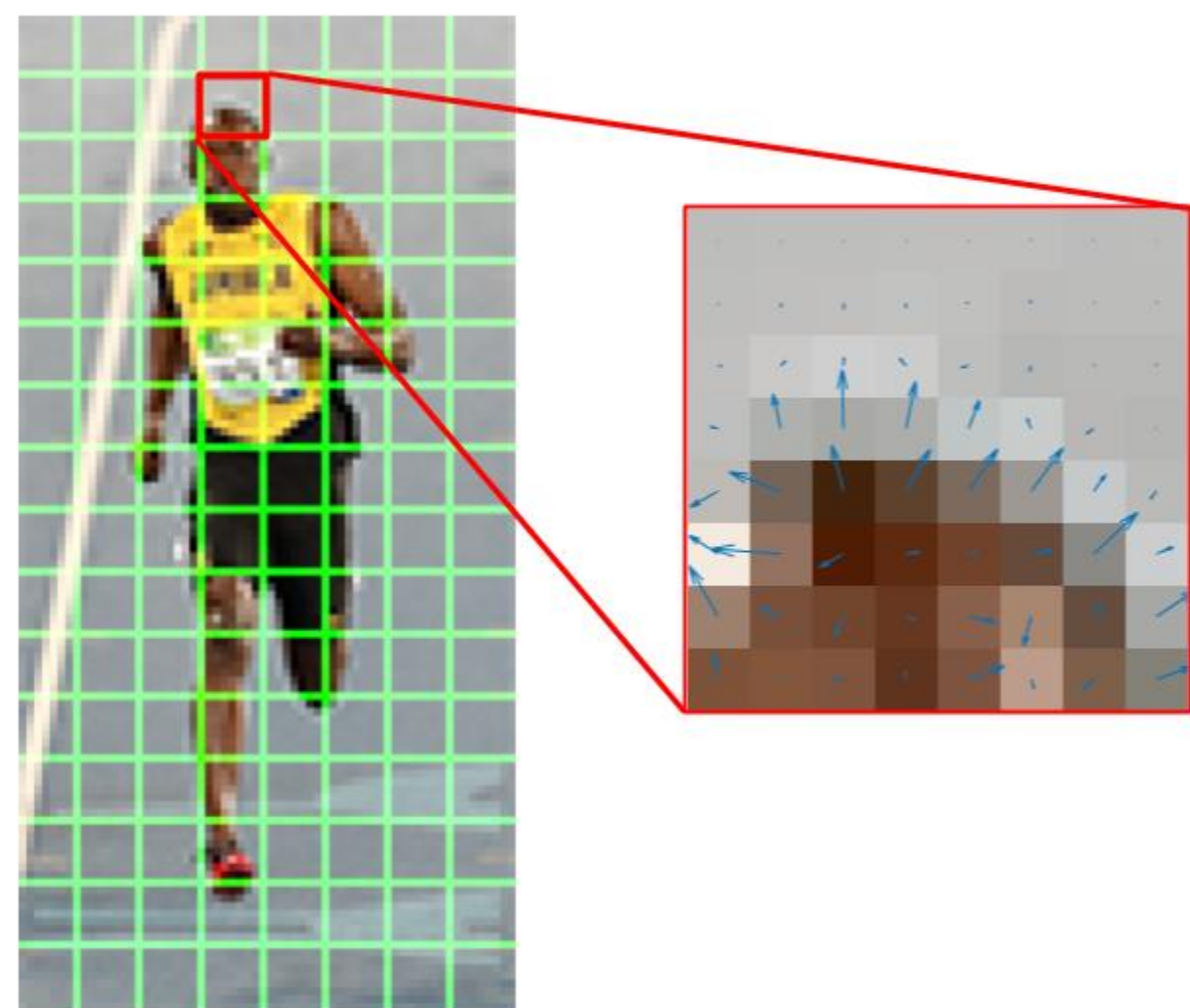
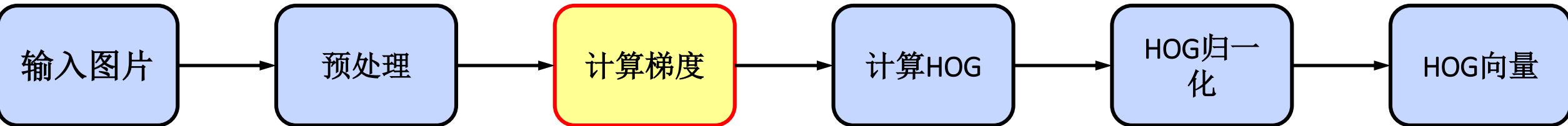


$$G_x = 240 - 190 = 50$$

$$G_y = 110 - 60 = 50$$

$$G = \sqrt{(G_x)^2 + (G_y)^2} = 50\sqrt{2}$$

$$\theta = \text{atan}\left(\frac{G_y}{G_x}\right) = 45^\circ$$



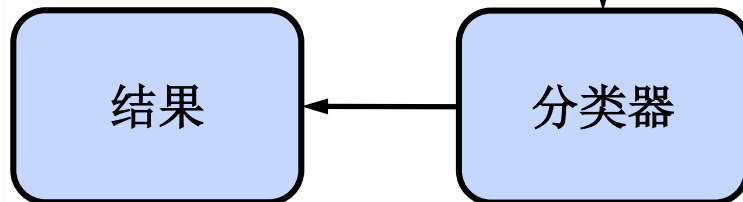
16X8=128 8x8 cells

2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

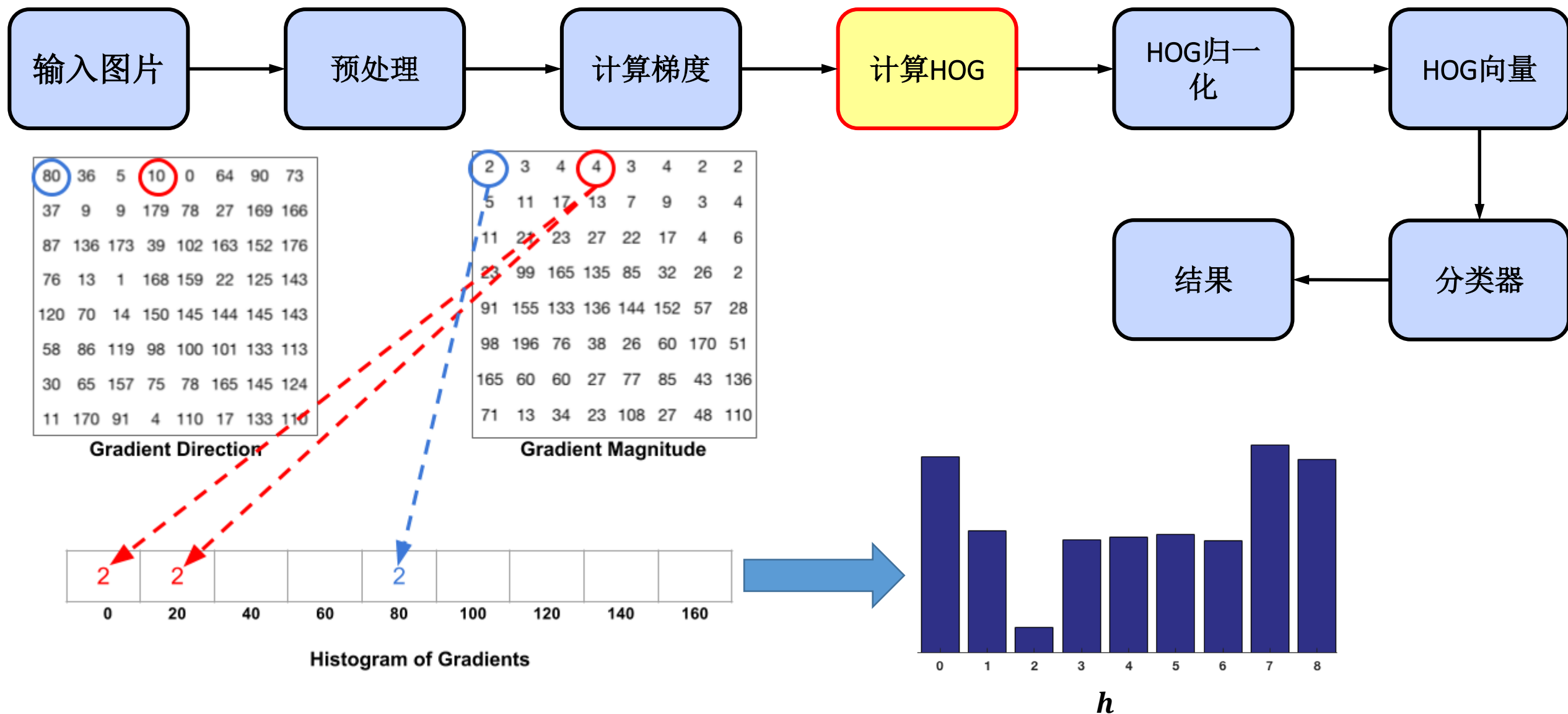
Gradient Magnitude

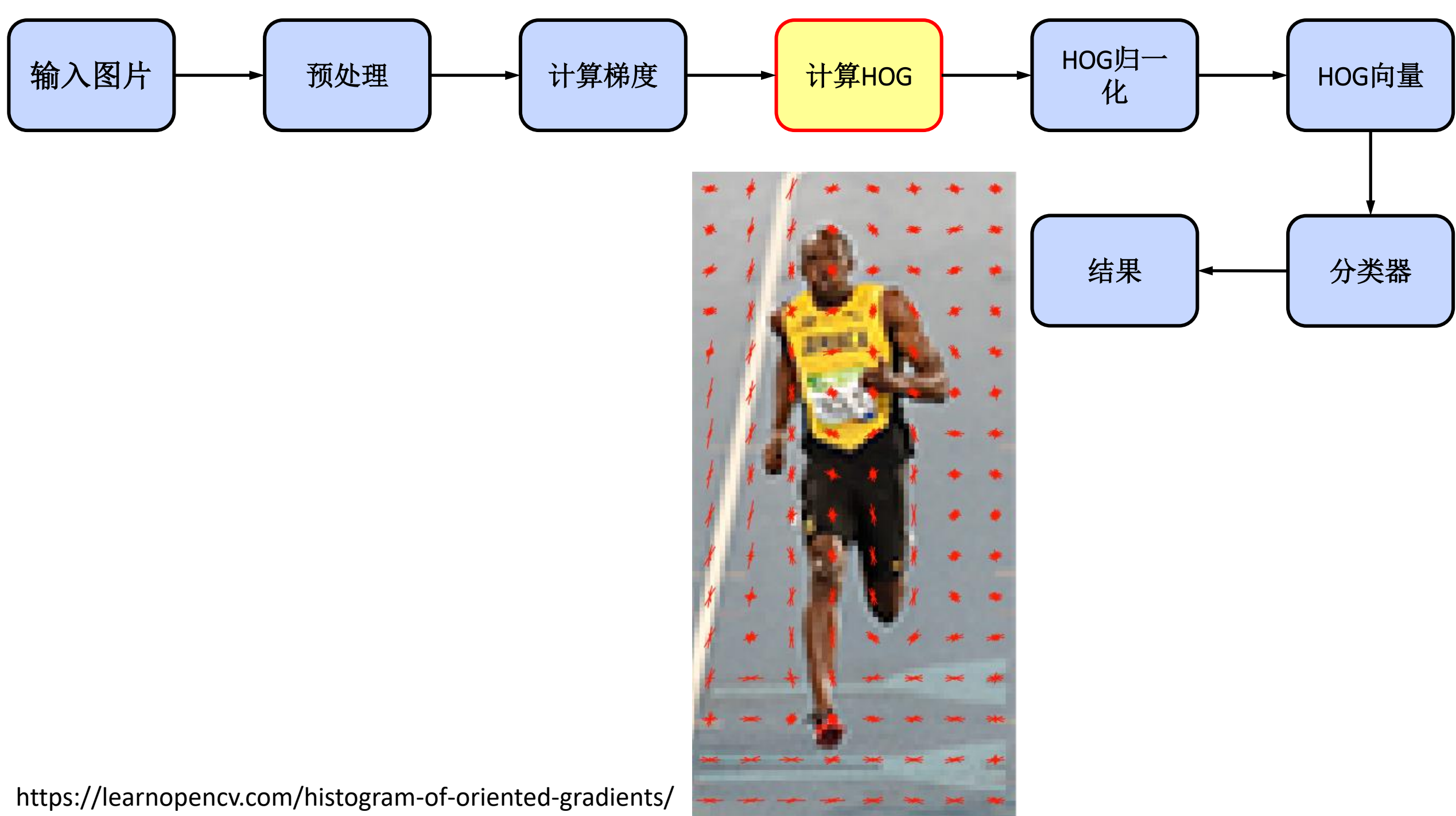
80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

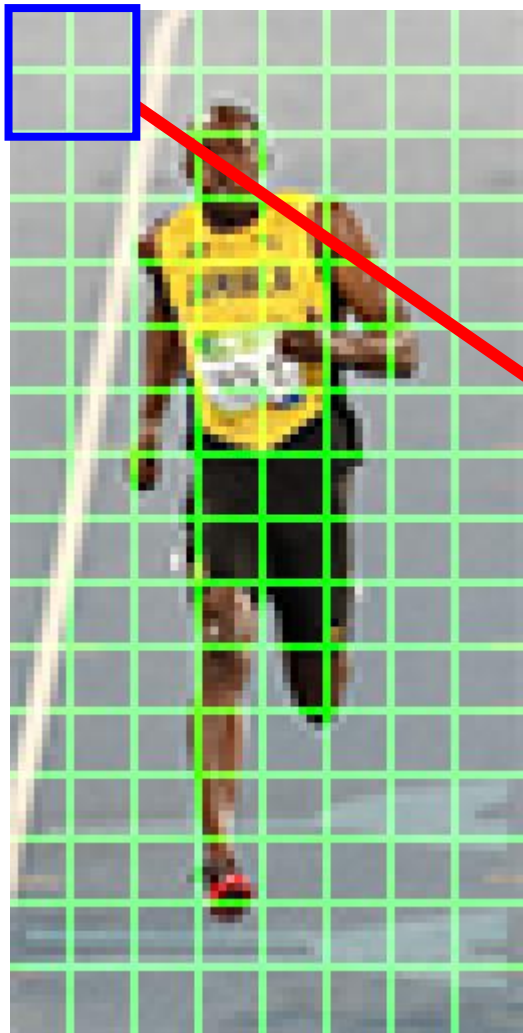
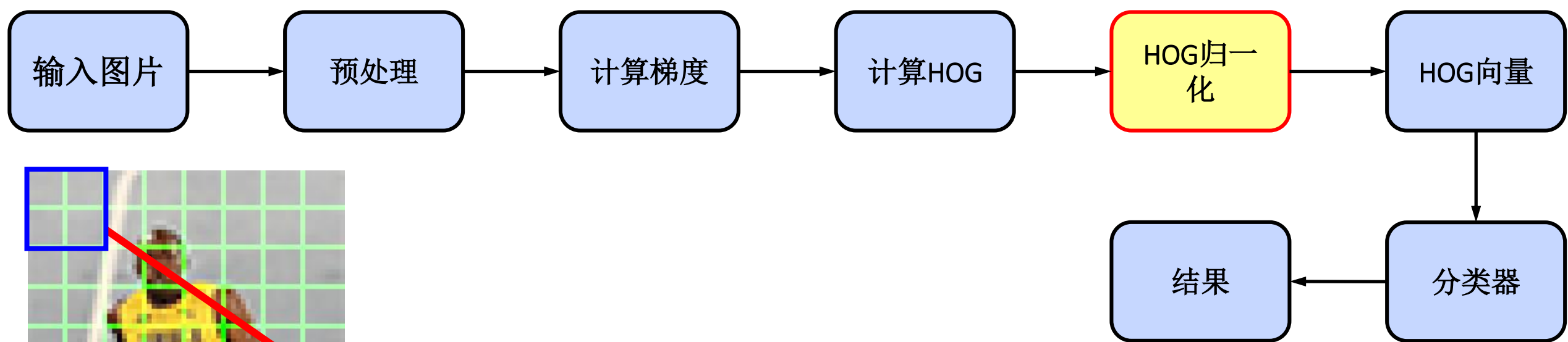
Gradient Direction



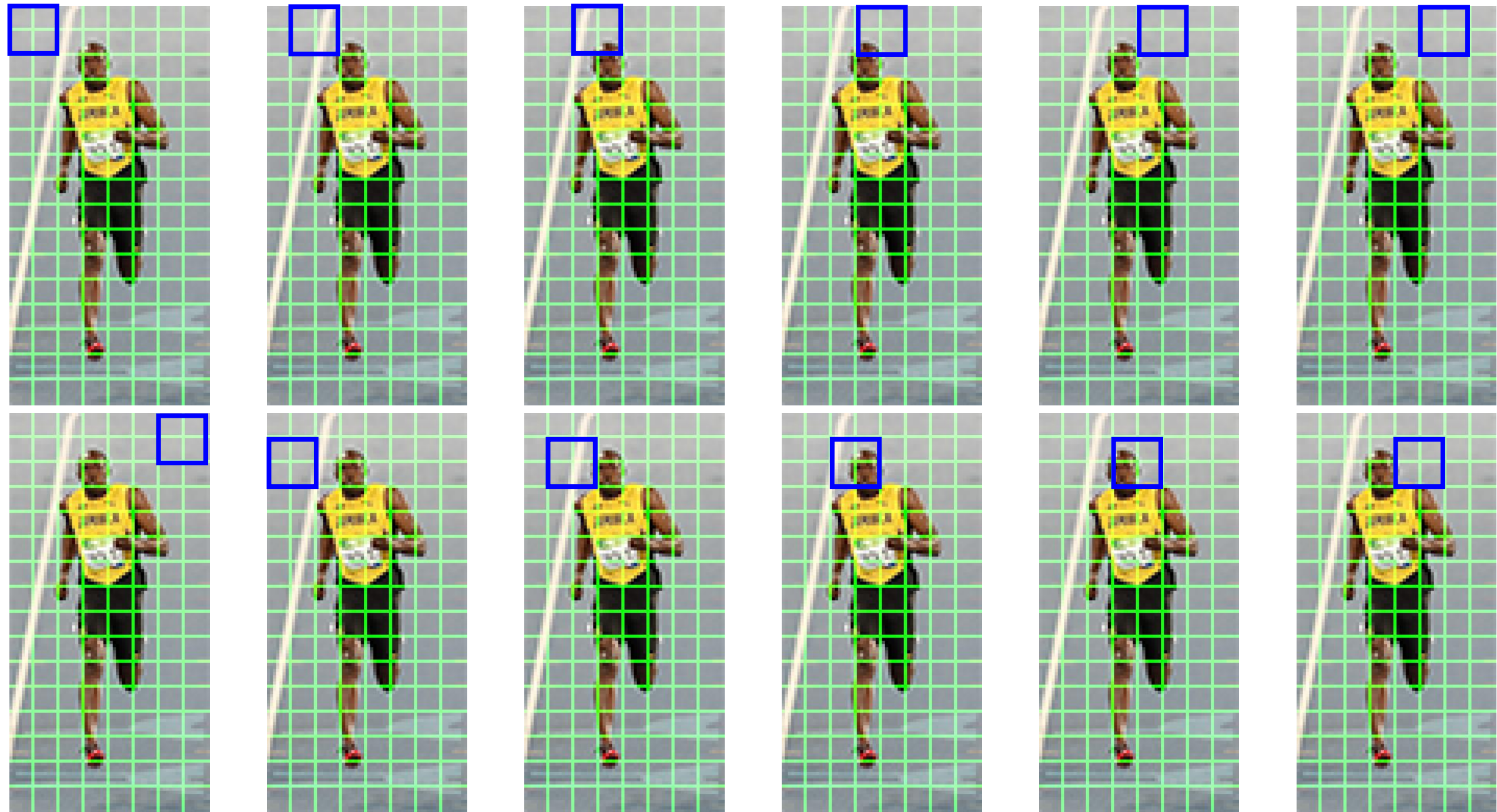
特征工程

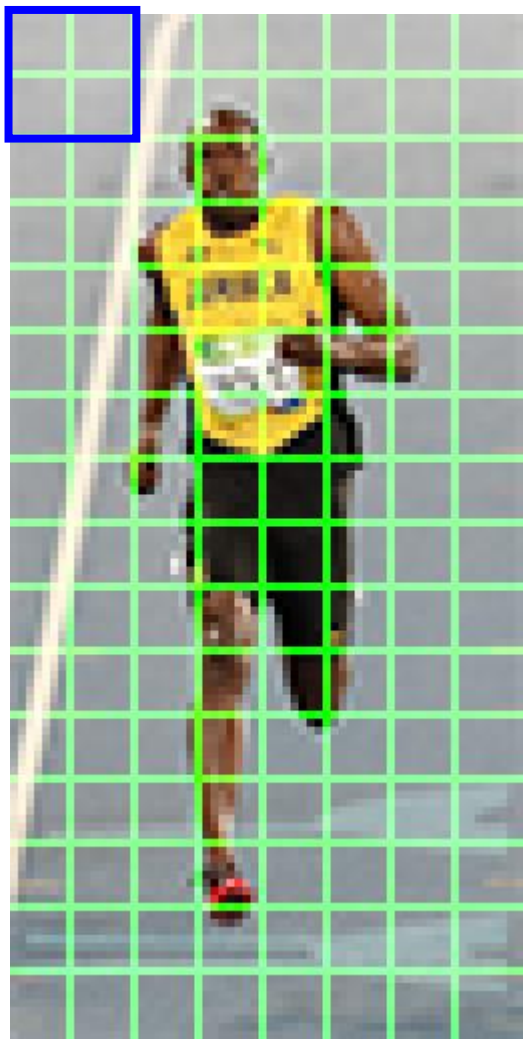
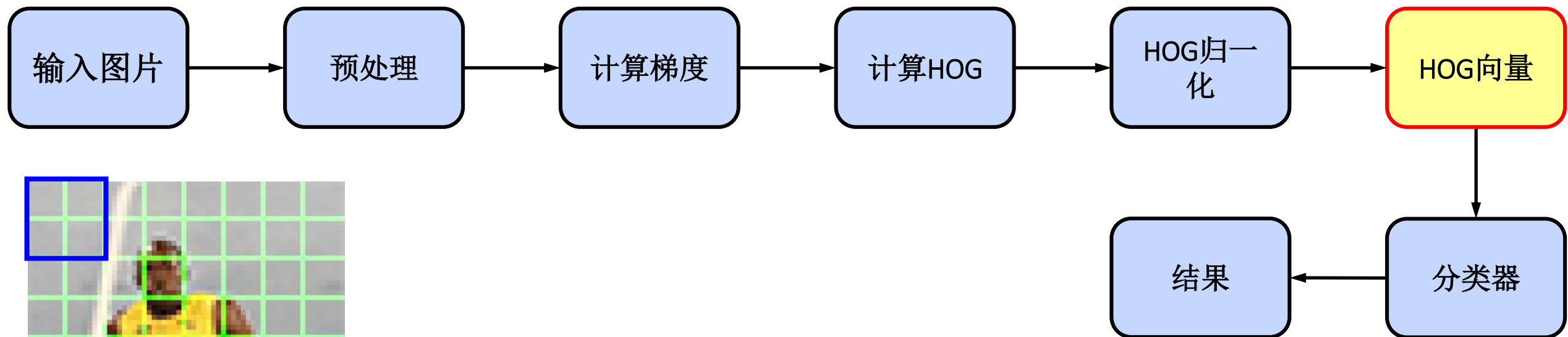






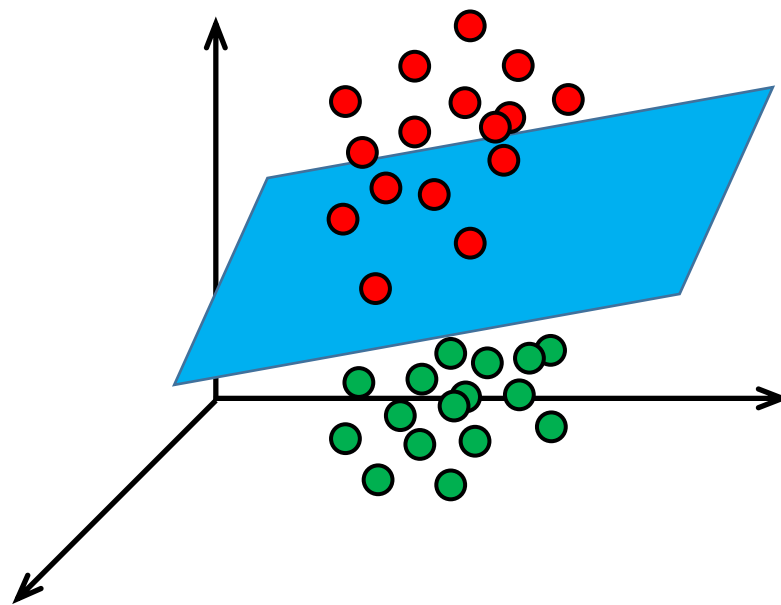
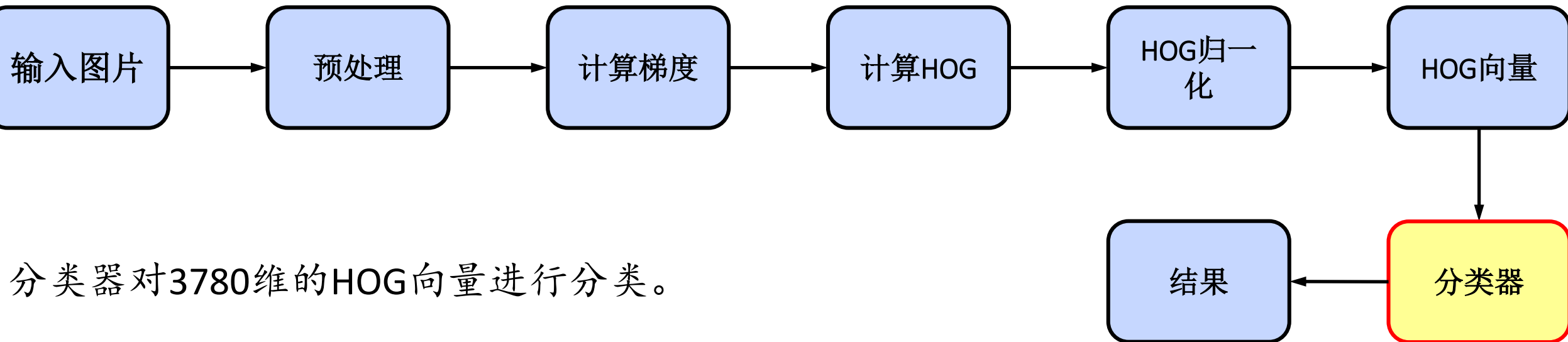
归一化具有36个元素的向量 $[h_{11} \quad h_{12} \quad h_{21} \quad h_{22}]$





- $7 \times 15 = 105$ 窗口位置。
- 每个窗口给出 36×1 向量。
- 将这105个向量拼接起来，可得到一个具有 $105 \times 36 = 3780$ 个元素的向量。


特征工程



特征工程

词典={爱(1), 机(2), 器(3), 我(4), 习(5), 学(6)}

我	爱	机	器	学	习
4	1	2	3	6	5

“我爱机器学习”  [4 1 2 3 6 5]

特征工程

词典={爱(1), 机(2), 器(3), 我(4), 习(5), 学(6)}

我	爱	机	器	学	习
$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$

- 这种特征表示方式被称为独热编码（one-hot vector）
- 每个词的独热编码向量长度与词典中字的个数一样。如果词典很大，独热编码向量很长。
- 独热编码向量绝大多数元素为0。

特征工程

我	爱	机	器	学	习
$\begin{bmatrix} 0.1 \\ 0.9 \\ 0.2 \\ -0.4 \\ -0.3 \\ 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.3 \\ -0.5 \\ 0.1 \\ 0.4 \\ 0.9 \\ -0.2 \end{bmatrix}$	$\begin{bmatrix} -0.1 \\ 0.6 \\ 0.8 \\ 0.2 \\ 0.5 \\ 0.1 \end{bmatrix}$	$\begin{bmatrix} 0.9 \\ 0.1 \\ -0.7 \\ 0.6 \\ 0.4 \\ -0.9 \end{bmatrix}$	$\begin{bmatrix} 0.2 \\ -0.5 \\ -0.9 \\ 0.6 \\ -0.7 \\ 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.6 \\ 0.4 \\ -0.8 \\ 0.9 \\ 0.2 \\ 0.7 \end{bmatrix}$

- 这种特征表示方式被称嵌入编码（embedding vectors）
- 嵌入词向量绝大多数元素非零，因此也被称为分布式词向量（distributed word embeddings）
- 一般使用海量文本训练神经网络模型得到嵌入词向量。
- 嵌入词向量长度可控。

$V(\text{king}) - V(\text{queen}) \approx V(\text{man}) - V(\text{woman})$

$V(\text{father}) - V(\text{mother}) \approx V(\text{man}) - V(\text{woman})$

$V(\text{uncle}) - V(\text{aunt}) \approx V(\text{man}) - V(\text{woman})$

$V(\text{China}) - V(\text{France}) \approx V(\text{Beijing}) - V(\text{Paris})$

$V(\text{doctor}) - V(\text{nurse}) \approx V(\text{man}) - V(\text{woman})$

$V(\text{pilot}) - V(\text{flight attendant}) \approx V(\text{man}) - V(\text{woman})$

The doctor asked the nurse to wash his hands

The doctor asked the nurse to wash her hands.

大夫让护士把他的帽子带好。

大夫让护士把她的帽子带好。

- **特征选择**(feature selection)
 - 滤波器法(filter method): 仅利用数据, 不利用分类器。
 - 包装法(wrapper method): 选择特征子集, 训练分类器。
 - 嵌入法(embedding method): 更改损失函数, 使其具有特征选择性。
- **特征提取**(feature extraction)
 - 从已有特征创造出新的特征

模型评价

TP: true positive

- 真实类别为正，预测类别均为正.

FP: false positive

- 真实类别为负，预测类别为正

FN: false negative

- 真实类别为正，预测类别为负

TN: true negative

- 真实类别为负，预测类别为证

		真实类别	
		正	负
预测类别	正	TP	FP
	负	FN	TN

二分类问题的混淆矩阵(confusion matrix)

模型评价

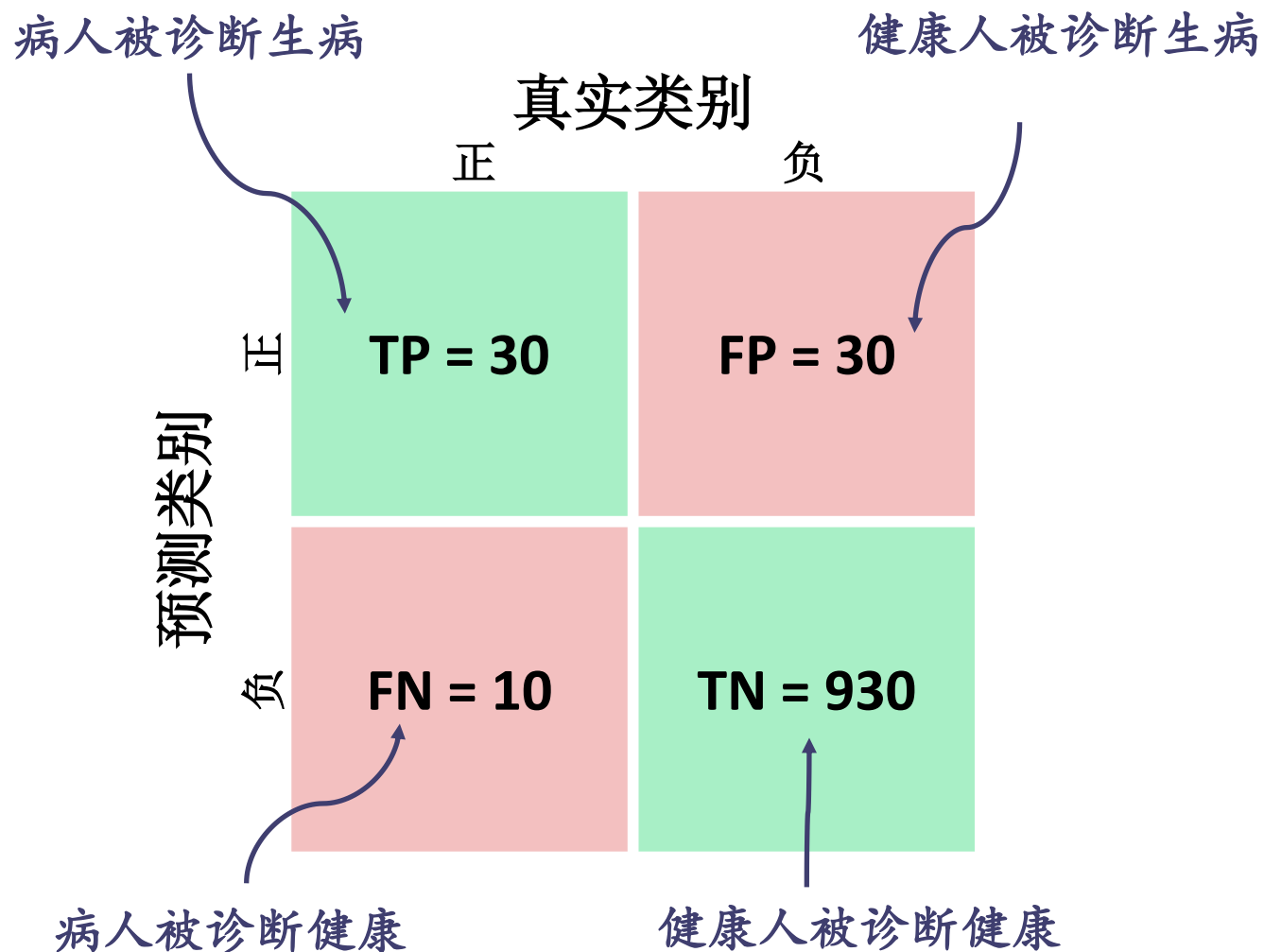
$$\text{准确率(accuracy)} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{召回率(recall or sensitivity)} = \frac{TP}{TP+FN}$$

$$\text{特异度(specificity)} = \frac{TN}{TN+FP}$$

$$\text{精度(precision)} = \frac{TP}{TP+FP}$$

$$F1\text{值} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



模型评价

ROC (线下区域, Region Under the Curve) 与AOC (线下区域面积, Area Under the Curve)

