

04 Naïve Bayesian Classifier

(朴素贝叶斯分类器)

本章目录

2

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

1.贝叶斯方法

3

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

1. 贝叶斯方法-背景知识

4

贝叶斯分类 : 贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。

先验概率 : 根据以往经验和分析得到的概率。我们用 $P(Y)$ 来代表在没有训练数据前假设 Y 拥有的初始概率。

后验概率 : 根据已经发生的事件来分析得到的概率。以 $P(Y|X)$ 代表假设 X 成立的情下观察到 Y 数据的概率，因为它反映了在看到训练数据 X 后 Y 成立的置信度。

1.贝叶斯方法-背景知识

5

联合概率： 联合概率是指在多元的概率分布中多个随机变量分别满足各自条件的概率。 X 与 Y 的联合概率表示为 $P(X, Y)$ 、 $P(XY)$ 或 $P(X \cap Y)$ 。

假设 X 和 Y 都服从正态分布，那么 $P(X < 5, Y < 0)$ 就是一个联合概率，表示 $X < 5, Y < 0$ 两个条件同时成立的概率。表示两个事件共同发生的概率。

1. 贝叶斯方法

6

贝叶斯公式

Diagram illustrating the components of Bayes' formula:

- 后验概率 (Posterior Probability) points to $P(Y|X)$
- 似然度 (Likelihood) points to $P(X|Y)$
- 先验概率 (Prior Probability) points to $P(Y)$
- 边际似然度 (Marginal Likelihood) points to $P(X)$

$$P(Y|X) = \frac{P(X, Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

朴素贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布 $P(X, Y)$ ，然后求得后验概率分布 $P(Y|X)$ 。

具体来说，利用训练数据学习 $P(X|Y)$ 和 $P(Y)$ 的估计，得到联合概率分布：

$$P(X, Y) = P(X|Y) P(Y)$$

1.贝叶斯方法

7

吸毒者检测

假设一个常规的检测结果的敏感度与可靠度均为99%，也就是说，当被检者吸毒时，每次检测呈阳性（+）的机率为99%。而被检者不吸毒时，每次检测呈阴性（-）的机率为99%。从检测结果的机率来看，检测结果是比较准确的，但是贝叶斯定理却可以揭示一个潜在的问题。假设某公司将对其全体雇员进行一次鸦片吸食情况的检测，已知0.5%的雇员吸毒。我们想知道，每位医学检测呈阳性的雇员吸毒的机率有多高？

1.贝叶斯方法

8

令「D」为雇员吸毒事件，「N」为雇员不吸毒事件，「+」为检测呈阳性事件。 可得：

1. $P(D)$ 代表雇员吸毒的机率，不考虑其他情况，该值为0.005。 因为公司的预先统计表明该公司的雇员中有0.5%的人吸食毒品，所以这个值就是D的[先验机率](#)。
2. $P(N)$ 代表雇员不吸毒的机率，显然，该值为0.995，也就是 $1-P(D)$ 。
3. $P(+|D)$ 代表吸毒者阳性检出率，这是一个[条件机率](#)，由于阳性检测准确性是99%，因此该值为0.99。
4. $P(+|N)$ 代表不吸毒者阳性检出率，也就是出错检测的机率，该值为0.01，因为对于不吸毒者，其检测为阴性的机率为99%，因此，其被误检测成阳性的机率为1-99%。
5. $P(+)$ 代表不考虑其他因素的影响的阳性检出率。 该值为0.0149或者1.49%。 我们可以通过全机率公式计算得到：此机率 = 吸毒者阳性检出率 ($0.5\% \times 99\% = 0.495\%$) + 不吸毒者阳性检出率 ($99.5\% \times 1\% = 0.995\%$)。 $P(+)$ = 0.0149是检测呈阳性的[先验机率](#)。 用数学公式描述为：

$$P(+) = P(+, D) + P(+, N) = P(+|D)P(D) + P(+|N)P(N)$$

1.贝叶斯方法

9

根据上述描述，我们可以计算某人检测呈阳性时确实吸毒的条件机率 $P(D|+)$ ：

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+)} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|N)P(N)} \\ &= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\ &= 0.3322. \end{aligned}$$

尽管我们的检测结果可靠性很高，但是只能得出如下结论：如果某人检测呈阳性，那么此人是吸毒的机率只有大约33%，也就是说此人不吸毒的可能性比较大。

我们测试的条件（本例中指D，雇员吸毒）越难发生，发生误判的可能性越大。

2.朴素贝叶斯原理

10

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

2.朴素贝叶斯原理

11

判别模型和生成模型

监督学习方法又分

生成方法（ Generative approach ）和**判别方法**（ Discriminative approach ）

所学到的模型分别称为

生成模型（ Generative Model ）和**判别模型**（ Discriminative Model ）。

判别模型（ Discriminative Model ）	生成模型（ Generative Model ）
由数据直接学习决策函数 $Y=f(X)$ 或者条件概率分布 $P(Y X)$ 作为预测的模型，即判别模型。基本思想是有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计 $P(Y X)$	由训练数据学习联合概率分布 $P(X, Y)$ ，然后求得后验概率分布 $P(Y X)$ 。具体来说，利用训练数据学习 $P(X Y)$ 和 $P(Y)$ 的估计，得到联合概率分布 $P(X, Y) = P(Y)P(X Y)$ ，再利用它进行分类。 即：估计 $P(X Y)$ 然后推导 $P(Y X)$
线性回归、逻辑回归、感知机、决策树、支持向量机.....	朴素贝叶斯、HMM、深度信念网络(DBN).....

2.朴素贝叶斯原理

12

1 . 朴素贝叶斯法是典型的生成学习方法。

生成方法由训练数据学习联合概率分布 $P(X, Y)$, 然后求得后验概率分布 $P(Y|X)$ 。具体来说 , 利用训练数据学习 $P(X|Y)$ 和 $P(Y)$ 的估计 , 得到联合概率分布 :

$$P(X, Y) = P(Y)P(X|Y)$$

概率估计方法可以是极大似然估计或贝叶斯估计。

2.朴素贝叶斯原理

13

2 . 朴素贝叶斯法的基本假设是条件独立性。

$$P(X = x|Y = c_k) = P(x^{(1)}, \dots, x^{(d)}|y^k) = \prod_{i=1}^d P(x^{(i)}|Y = c_k)$$

c_k 代表类别， k 代表类别个数。

这是一个较强的假设。由于这一假设，模型包含的条件概率的数量大为减少，朴素贝叶斯法的学习与预测大为简化。因而朴素贝叶斯法高效，且易于实现。其缺点是分类的性能不一定很高。

2.朴素贝叶斯原理

14

3 . 朴素贝叶斯法利用贝叶斯定理与学到的联合概率模型进行分类预测

我们要求的是 $P(Y|X)$ ，根据生成模型定义我们可以求 $P(X, Y)$ 和 $P(Y)$ 假设中的特征是条件独立的。这个称作朴素贝叶斯假设。形式化表示为，（如果给定 Z 的情况下， X 和 Y 条件独立）：

$$P(X|Z) = P(X|Y, Z)$$

也可以表示为：

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

2.朴素贝叶斯原理

15

用于文本分类的朴素贝叶斯模型，这个模型称作多值伯努利事件模型。

在这个模型中，我们首先随机选定了邮件的类型 $p(y)$ ，然后一个人翻阅词典的所有词，随机决定一个词是否出现依照概率 $p(x^{(1)}|y)$ ，出现标示为1，否则标示为0。假设有50000个单词，那么这封邮件的概率可以表示为：

$$\begin{aligned} & p(x^{(1)}, \dots, x^{(50000)}|y) \\ &= p(x^{(1)}|y)p(x^{(2)}|y, x^{(1)})p(x^{(3)}|y, x^{(1)}, x^{(2)}) \cdots p(x^{(50000)}|y, x^{(1)}, \dots, x^{(49999)}) \\ &= p(x^{(1)}|y)p(x^{(2)}|y)p(x^{(3)}|y) \cdots p(x^{(50000)}|y) \\ &= \prod_{i=1}^m p(x^{(i)}|y) \end{aligned}$$

2.朴素贝叶斯原理

16

独立性

将输入 x 分到后验概率最大的类 y 。

$$y = \operatorname{argmax}_{c_k} P(Y = c_k) \prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k)$$

2.朴素贝叶斯原理

17

$$y = \operatorname{argmax}_{c_k} P(Y = c_k) \prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k)$$

$X = \{X^1, \dots, X^n\}$ ，为 n 维向量的集合

$Y = \{Y^1, \dots, Y^n\}$ ， $Y^n \in \{c_1, c_2, \dots, c_k\}$ ， K 为类别数

训练数据集 $T = \{(X^1, Y^1), (X^2, Y^2), \dots, (X^n, Y^n)\}$ 由 $P(X, Y)$ 独立同分布产生。

2.朴素贝叶斯原理

18

朴素贝叶斯法对**条件概率分布作了条件独立性的假设**。由于这是一个较强的假设，朴素贝叶斯法也由此得名。具体地，条件独立性假设是：

$$\begin{aligned} P(X = x|Y = c_k) &= P(X_1 = x^{(1)}, X_2 = x^{(2)}, \dots, X_d = x^{(d)}|Y = c_k) \\ &= \prod_{i=1}^d P(X_i = x^{(i)}|Y = c_k) \end{aligned} \tag{1}$$

2.朴素贝叶斯原理

19

朴素贝叶斯法分类时，对给定的输入 x ，通过学习到的模型计算后验概率分布 $P(Y = c_k|X = x)$ ，将后验概率最大的类作为 x 的类输出。根据贝叶斯定理：

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

可以计算后验概率

$$P(Y = c_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_{k=1}^K P(X = x|Y = c_k)P(Y = c_k)} \quad (2)$$

2.朴素贝叶斯原理

20

将式(1)代入公式(2)，可以得到

$$P(Y = c_k | X = x) = \frac{\prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k) P(Y = c_k)}{\sum_{k=1}^K \prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k) P(Y = c_k)}$$

贝叶斯分类器可以表示为：

$$y = f(x) = \operatorname{argmax}_{c_k} \frac{\prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k) P(Y = c_k)}{\sum_{k=1}^K \prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k) P(Y = c_k)}$$

上式中分母中 c_k 都是一样的，即不会对结果产生影响，即

$$y = f(x) = \operatorname{argmax}_{c_k} \prod_{i=1}^d P(X_i = x^{(i)} | Y = c_k) P(Y = c_k)$$


2.朴素贝叶斯原理

21

Task: Predict whether or not a picnic spot is enjoyable

Training Data: $X = (X_1 \quad X_2 \quad X_3 \quad \dots \quad \dots \quad X_d) \quad Y$

n rows



Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Naïve Bayes assumption: $P(X_1 \dots X_d | Y) = \prod_{i=1}^d P(X_i | Y)$

How many parameters to estimate?

(X is composed of d binary features, Y has K possible class labels)

$(2^d - 1)K$ vs $(2 - 1)dK$

2.朴素贝叶斯原理

22

Given:

- Class prior $P(Y)$
- d conditionally independent features X_1, \dots, X_d given the class label Y
- For each X_i feature, we have the conditional likelihood $P(X_i|Y)$

Naïve Bayes Decision rule:

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg \max_y P(x_1, \dots, x_d | y) P(y) \\ &= \arg \max_y \prod_{i=1}^d P(x_i | y) P(y) \end{aligned}$$

2.朴素贝叶斯原理

23


Training data: $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$

$$X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$$

n d -dimensional discrete features + K class labels

$$f_{NB}(\mathbf{x}) = \arg \max_y \prod_{i=1}^d P(x_i|y)P(y)$$

We need to estimate these probabilities!



Estimate them with MLE (Relative Frequencies)!

2.朴素贝叶斯原理

24

$$f_{NB}(\mathbf{x}) = \arg \max_y \prod_{i=1}^d P(x_i|y)P(y) \quad \text{We need to estimate these probabilities!}$$

Estimators

For Class Prior

$$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

For Likelihood

$$\frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

NB Prediction for test data:

$$X = (x_1, \dots, x_d)$$

$$Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

3.朴素贝叶斯案例

25

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

3.朴素贝叶斯案例-离散特征

26

假设我们正在构建一个分类器，该分类器说明文本是否与运动(Sports)有关。我们的训练数据有5句话：

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

我们想要计算句子“A very close game”是 Sports 的概率以及它不是 Sports 的概率。

即 $P(\text{Sports} \mid \text{a very close game})$ 这个句子的类别是Sports的概率

3.朴素贝叶斯案例-离散特征

27

特征：单词的频率

已知贝叶斯定理 $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$ ，则：

$$P(\text{Sports} | \text{a very close game}) = \frac{P(\text{a very close game} | \text{Sports}) \times P(\text{Sports})}{P(\text{a very close game})}$$

由于我们只是试图找出哪个类别有更大的概率，可以舍弃除数，只是比较

$P(\text{a very close game} | \text{Sports}) \times P(\text{Sports})$ 和

$P(\text{a very close game} | \text{Not Sports}) \times P(\text{Not Sports})$

3.朴素贝叶斯案例-离散特征

28

我们假设一个句子中的每个单词都与其他单词无关。

$$\begin{aligned} &P(\text{a very close game}) \\ &= P(a) \times P(\text{very}) \times P(\text{close}) \times P(\text{game}) \end{aligned}$$

$$\begin{aligned} &P(\text{a very close game} | \text{Sports}) \\ &= P(a | \text{Sports}) \times P(\text{very} | \text{Sports}) \times P(\text{close} | \text{Sports}) \times P(\text{game} | \text{Sports}) \end{aligned}$$

3.朴素贝叶斯案例-离散特征

29

计算每个类别的先验概率：

对于训练集中的给定句子，

$P(\text{Sports})$ 的概率为 $\frac{3}{5}$ 。

$P(\text{Not Sports})$ 是 $\frac{2}{5}$ 。

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

然后，在计算 $P(\text{game}|\text{Sports})$ 就是 “game” 有多少次出现在 Sports 的样本，然后除以 sports 为标签的文本的单词总数（ $3+3+5=11$ ）。

因此， $P(\text{game}|\text{Sports}) = \frac{2}{11}$ 。

“close” 不会出现在任何 sports 样本中！那就是说 $P(\text{close}|\text{Sports}) = 0$ 。

3.朴素贝叶斯案例-离散特征

30

通过使用一种称为**拉普拉斯平滑**的方法：我们为每个计数加1，因此它永远不会为零。为了平衡这一点，我们将可能单词的数量添加到除数中，因此计算结果永远不会大于1。



14个单词

在这里的情况下，可能单词是['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match']。

由于可能的单词数是14，因此应用平滑处理可以得到

$$P(\text{game} \mid \text{sports}) = \frac{2+1}{11+14}$$

3.朴素贝叶斯案例-离散特征

31

拉普拉斯平滑是一种用于平滑分类数据的技术。引入拉普拉斯平滑法来解决零概率问题,通过应用此方法,先验概率和条件概率可以写为

$$P_{\lambda}(C_k) = P_{\lambda}(Y = C_k) = \frac{\sum_{j=1}^N I(y_j = C_k) + \lambda}{N + K\lambda}$$

$$P_{\lambda}(x_j = a_i | y = C_k) = \frac{\sum_{j=1}^N I(x_j = a_i, y_j = C_k) + \lambda}{\sum_{j=1}^N I(y_j = C_k) + A\lambda}$$

其中 K 表示类别数量, A 表示 a_i 中不同值的数量, 通常 $\lambda = 1$

加入拉普拉斯平滑之后, 避免了出现概率为0的情况, 又保证了每个值都在0到1的范围内, 又保证了最终和为1的概率性质。

3.朴素贝叶斯案例-离散特征

32

Word	P (word Sports)	P (word Not Sports)
a	$(2 + 1) \div (11 + 14)$	$(1 + 1) \div (9 + 14)$
very	$(1 + 1) \div (11 + 14)$	$(0 + 1) \div (9 + 14)$
close	$(0 + 1) \div (11 + 14)$	$(1 + 1) \div (9 + 14)$
game	$(2 + 1) \div (11 + 14)$	$(0 + 1) \div (9 + 14)$

$$P(a | \text{Sports}) \times P(\text{very} | \text{Sports}) \times P(\text{close} | \text{Sports}) \times P(\text{game} | \text{Sports}) \times P(\text{Sports}) \\ = 2.76 \times 10^{-5} = 0.0000276$$

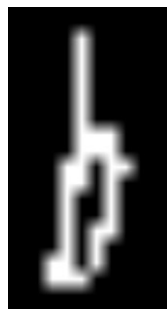
$$P(a | \text{Not Sports}) \times P(\text{very} | \text{Not Sports}) \times P(\text{close} | \text{Not Sports}) \\ \times P(\text{game} | \text{Not Sports}) \times P(\text{Not Sports}) \\ = 0.572 \times 10^{-5} = 0.00000572$$

由于0.0000276大于0.00000572，我们的分类器预测 “A very close game” 是Sport类。

3.朴素贝叶斯案例-连续特征

33

Eg., character recognition: X_i is intensity at i^{th} pixel



Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Different mean and variance for each class k and each pixel i .

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

3.朴素贝叶斯案例-连续特征

34

$$\begin{aligned}h_{NB}(\mathbf{x}) &= \arg \max_y P(y) \prod_i P(X_i = x_i | y) \\ &\approx \arg \max_k \hat{P}(Y = k) \prod_i \mathcal{N}(\hat{\mu}_{ik}, \hat{\sigma}_{ik})\end{aligned}$$

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^N x_j$$

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \hat{\mu})^2$$

3.朴素贝叶斯案例-连续特征

35

Maximum likelihood estimates:

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^N x_j$$

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

i^{th} pixel in
 j^{th} training image

k^{th} class

j^{th} training image

$$\hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \hat{\mu})^2$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

3.朴素贝叶斯案例-连续特征

36

用于分类精神状态的Gaussian NB

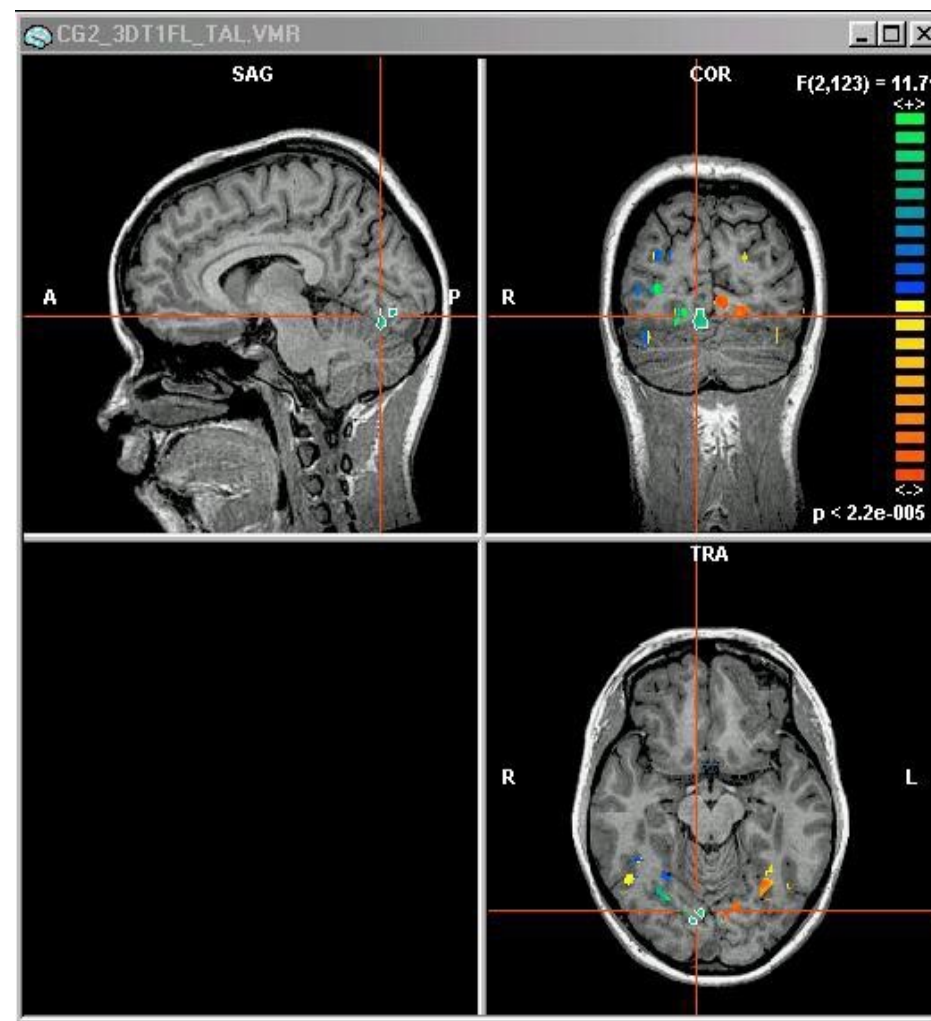


~1 mm resolution

~2 images per sec. 15,000

voxels/image non-invasive, safe

measures Blood Oxygen Level
Dependent (BOLD) response



[Mitchell et al.]

- [1] TOM M MICHELLE. Machine Learning[M]. New York: McGraw-Hill Companies, Inc, 1997.
- [2] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning[M]. New York: Springer, 2001.
- [3] CHRISTOPHER M. BISHOP. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006.
- [4] Zhang H., The optimality of naïve Bayes[C]//Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS), Miami, FL, 562-567, 2004.
- [5] Ng A. Y. and M. I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes[C]// Proceedings of the Advances in 14th Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 841-848, 2002.
- [6] Kohavi R., Scaling up the accuracy of naïve Bayes classifiers: A decision-tree hybrid[C]// Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, 202-207, 1996.
- [7] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2019.