生成式学习、高斯判别分析、朴素贝叶斯

Fei Gao

gaofei@hdu.edu.cn

https://fei-hdu.github.io/

杭州电子科技大学
HANGZHOU DIANZI UNIVERSITY

笃学力行 守正求新

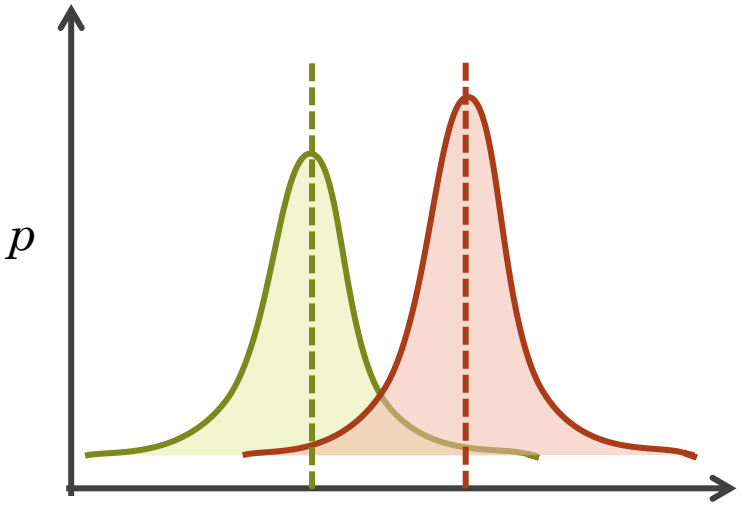
目录

1 Generative Learning algorithms

- **Discriminative learning algorithms**
  - Algorithms that try to learn $p(y|x)$ directly (such as logistic regression),
  - or algorithms that try to learn mappings directly from the space of inputs X to the labels:

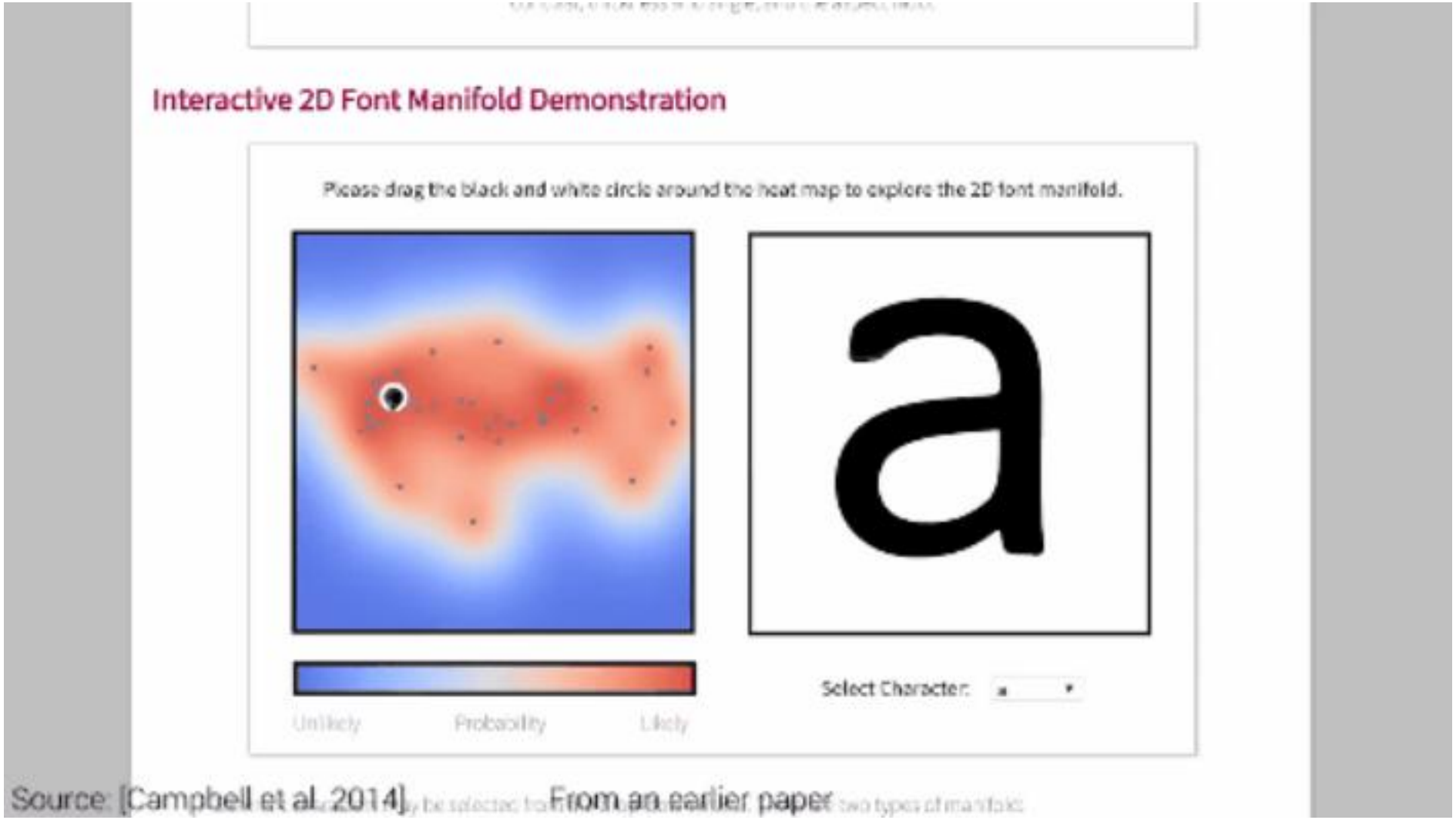
$$\mathcal{F} : \mathcal{X} \mapsto \{0, 1\}$$

  - (such as the perceptron algorithm)



- **Generative Learning algorithms**
  - algorithms that instead try to model $p(x|y)$ (and $p(y)$).
  - For instance,
    - if y indicates whether an example is a dog (0) or an elephant (1), then
      - $p(x|y = 0)$ models the distribution of dogs' features, and
      - $p(x|y = 1)$ models the distribution of elephants' features.

1 Generative Learning algorithms



Source: [Campbell et al. 2014]. From an earlier paper

1 Generative Learning algorithms

- **Bayes Rule**
  - After modeling p(y) (called the class priors) and p(x | y), our algorithm can then use Bayes rule to derive the posterior distribution on y given x:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

  - Here, the denominator is given by $p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$
  - Actually, if were calculating p(y | x) in order to make a prediction, then we don't actually need to calculate the denominator, since

$$\begin{aligned}\arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y).\end{aligned}$$

目录

2 Gaussian discriminant analysis (GDA)

• 2.1 The multivariate normal distribution

- **mean vector** $\mu \in \mathbb{R}^n$
- **covariance matrix** $\Sigma \in \mathbb{R}^{n \times n}$

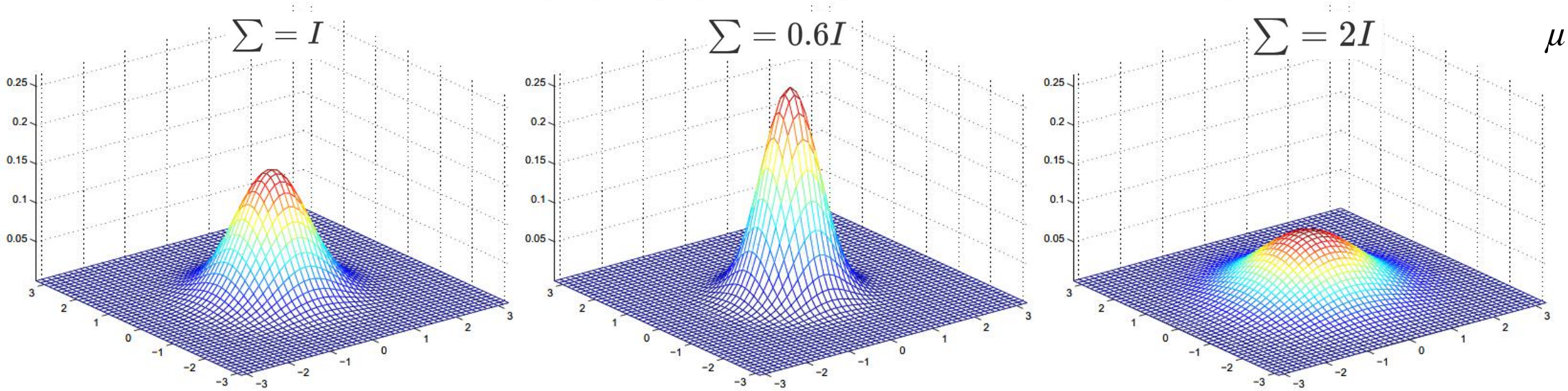
} "$\mathcal{N}(\mu, \Sigma)$"

σ

- symmetric and positive semi-definite matrix

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

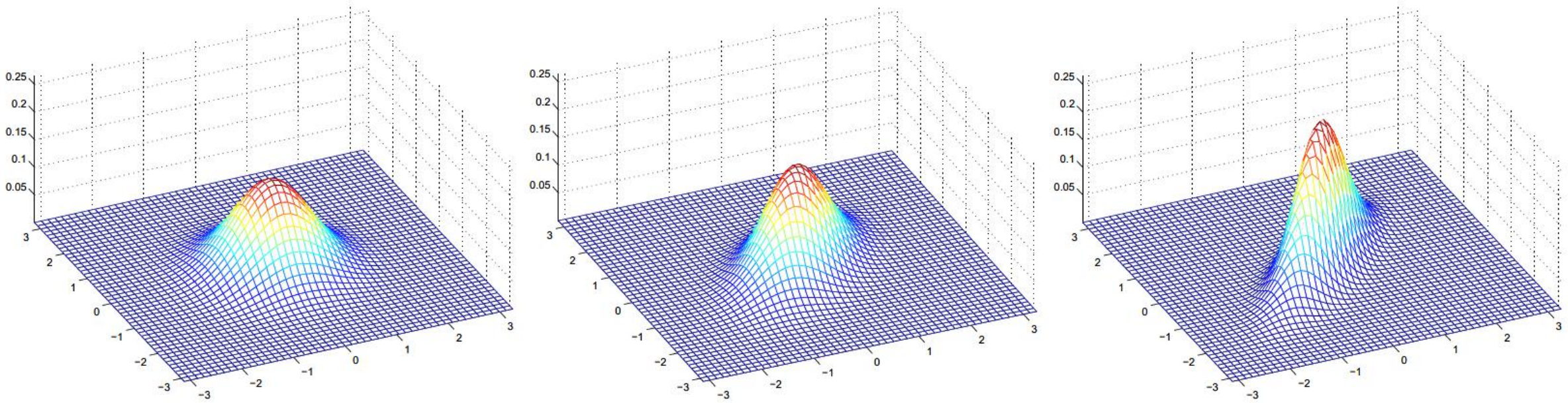
$\mu = 0$



2 Gaussian discriminant analysis (GDA)

- **2.1 The multivariate normal distribution**

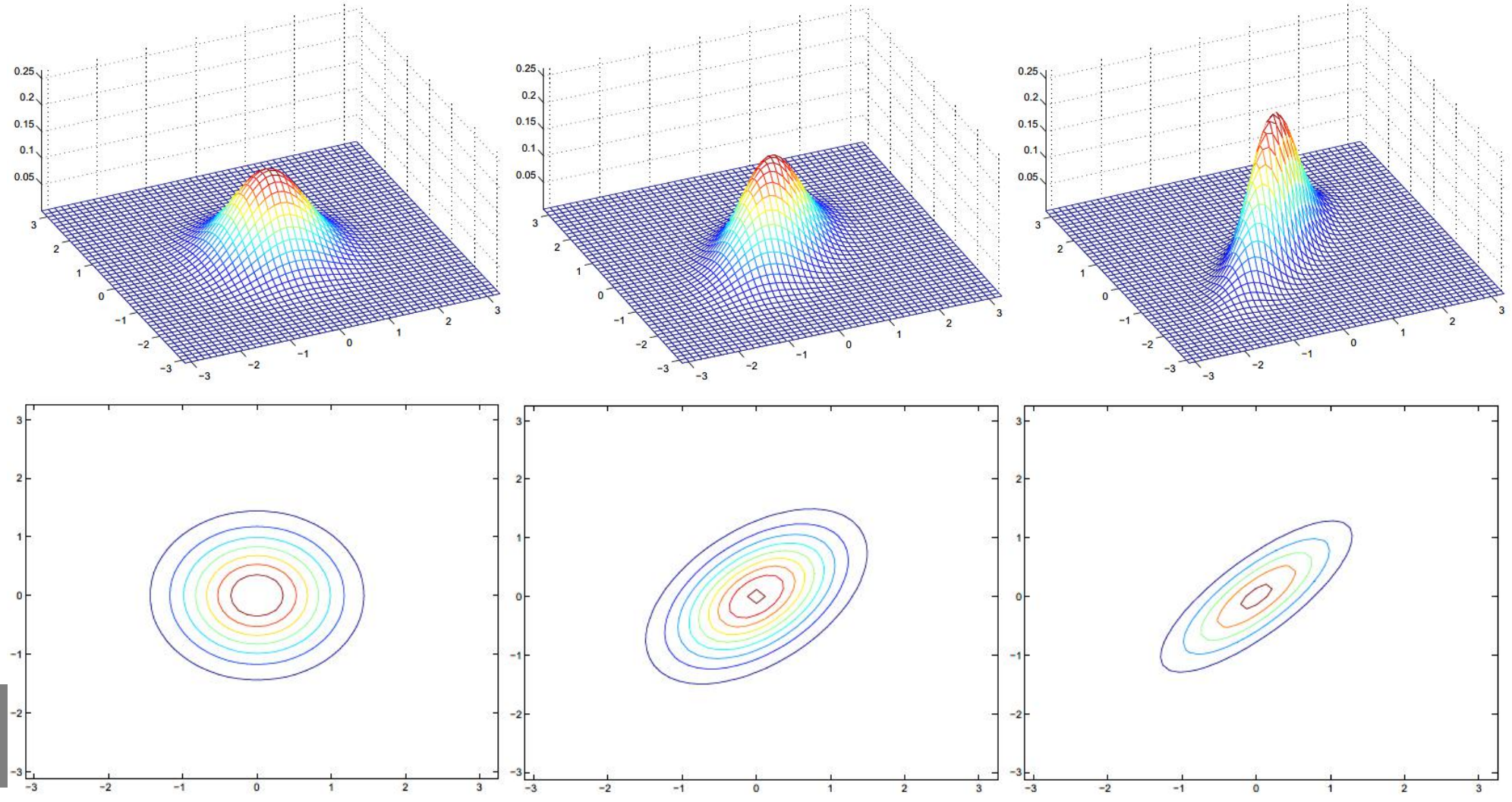
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}. \qquad \mu = 0$$



As we increase the off-diagonal entry in the covariance marix, the density becomes more “compressed” towards the 45◦ line.

2 Gaussian discriminant analysis (GDA)

- **2.1 The multivariate normal distribution**

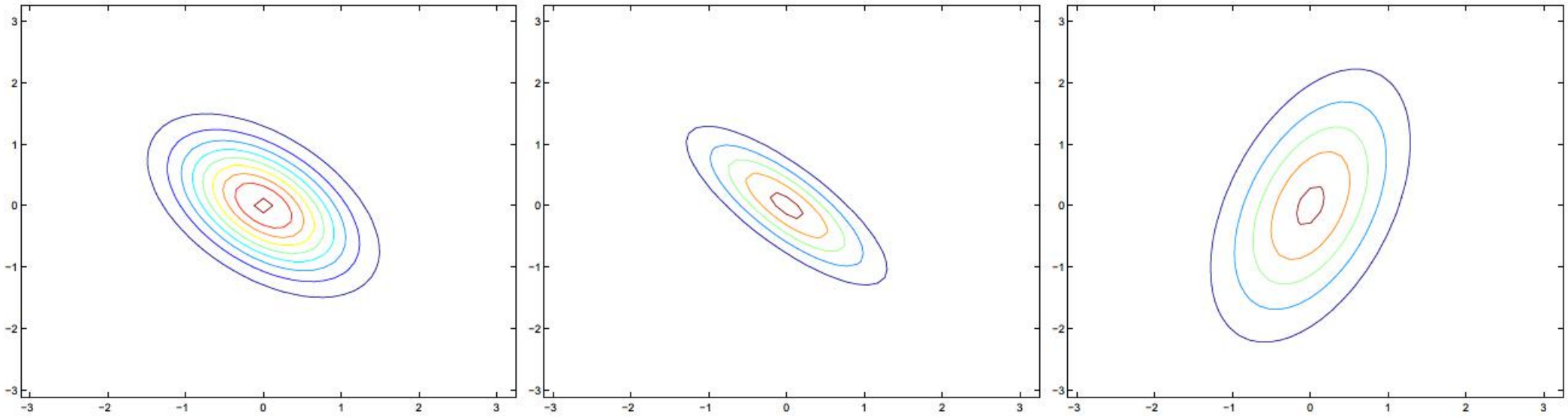


the contours of the same three densities

2 Gaussian discriminant analysis (GDA)

- **2.1 The multivariate normal distribution**

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}. \qquad \mu = 0$$

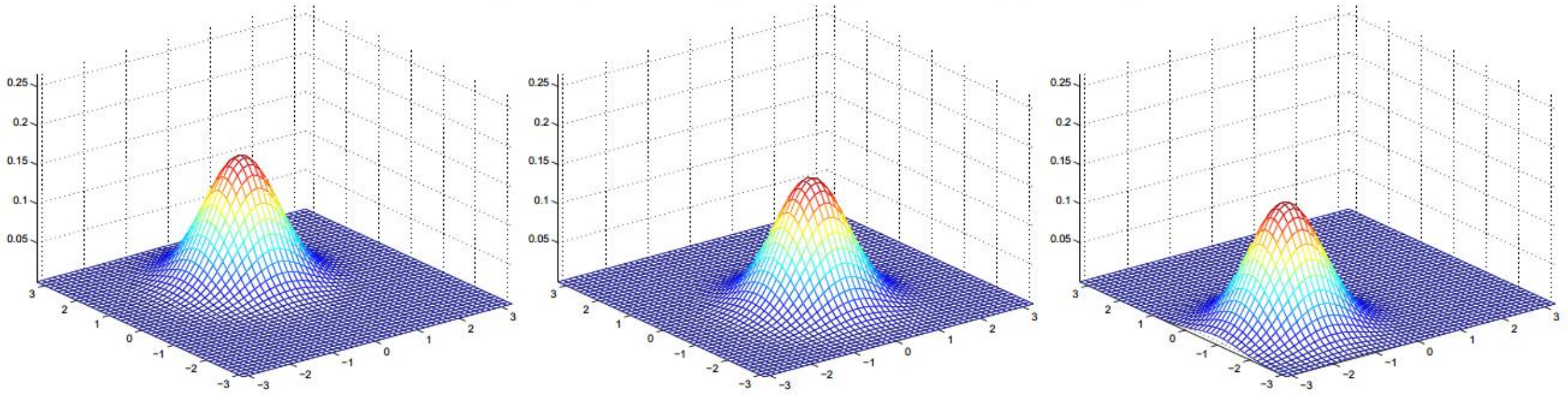


- By decreasing the off-diagonal elements of the covariance matrix, the density now becomes “compressed” again, but in the opposite direction.
- As we vary the parameters, more generally the contours will form ellipses .

2 Gaussian discriminant analysis (GDA)

- **2.1 The multivariate normal distribution**

$$\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}. \qquad \Sigma = I$$



- By varying μ, we can also move the mean of the density around.

2 Gaussian discriminant analysis (GDA)

- **2.2 The Gaussian Discriminant Analysis model**
  - GDA models $p(x|y)$ using a multivariate normal distribution. The model is:

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned}$$

  - Writing out the distributions, this is:

$$\begin{aligned} p(y) &= \phi^y(1-\phi)^{1-y} \\ p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) \end{aligned}$$

2 Gaussian discriminant analysis (GDA)

- **2.2 The Gaussian Discriminant Analysis model**
  - Here, the parameters of our model are ϕ, Σ, μ_0 and μ_1. (Note that while there're two different mean vectors μ_0 and μ_1, this model is usually applied using only one covariance matrix Σ.) The log-likelihood of the data is given by

$$\begin{aligned}\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{m} p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi).\end{aligned}$$

  - they share a covariance matrix, but they have different means

2 Gaussian discriminant analysis (GDA)

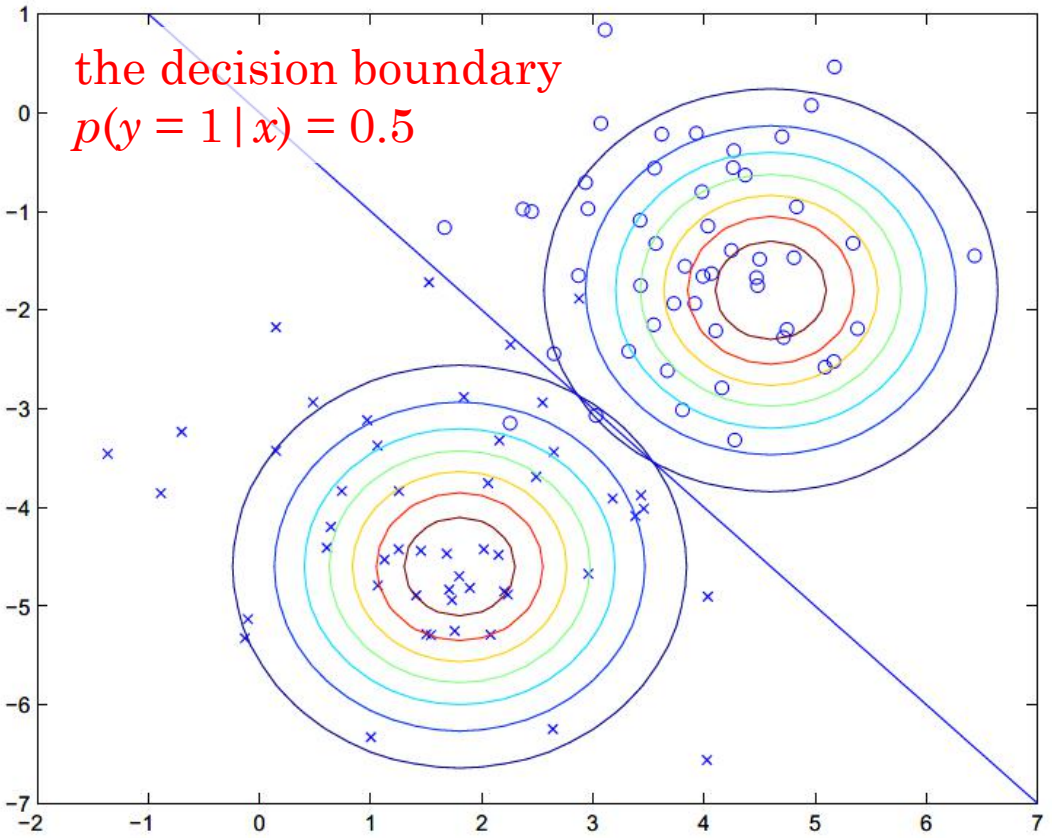
- **2.2 The Gaussian Discriminant Analysis model**
  - By maximizing ℓ with respect to the parameters, we find the maximum likelihood estimate of the parameters (see problem set 1) to be:

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$



目录

3 Naive Bayes（朴素贝叶斯）

- **托马斯·贝叶斯（Thomas Bayes）**

  - 英国“业余”数学家
  - 约1701年出生于伦敦，做过神甫
    - 统计学家认为概率这个东西就是上帝在掷骰子。
  - 1742年成为英国皇家学会会员。
  - 1761年4月7日逝世。
  - 1763年发表的一篇论文

Recall: Generative Learning algorithms

- **Generative Learning algorithms**
  - algorithms that instead try to model $p(x \mid y)$ (and $p(y)$).
  - For instance,
    - if y indicates whether an example is a dog (0) or an elephant (1), then
      - $p(x \mid y = 0)$ models the distribution of dogs' features, and
      - $p(x \mid y = 1)$ models the distribution of elephants' features.

- **Bayes Rule**
  - use Bayes rule to derive the posterior distribution on y given x:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \qquad P(y \mid x) = \frac{P(x,y)}{P(x)} = \frac{P(x \mid y)P(y)}{P(x)}$$

  - Here, the denominator is given by $p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$

3 Naive Bayes

- **3.1 Example1: Sparm E-mail Classification (text classification)**

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix} \quad \mapsto \quad y$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- **Naive Bayes (NB) assumption**
  - the x_i's are conditionally independent given y.
- **Naive Bayes classifier**
  - the resulting algorithm is called the Naive Bayes classifier.

3 Naive Bayes

• 3.1 Example1: Sparm E-mail Classification (text classification)

- if we have, say, a vocabulary of 50000 words, then

$$x \in \{0,1\}^{50000}$$

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix} \quad \mapsto \quad y$$

- We have

$$\begin{aligned} & p(x_1, \ldots, x_{50000}|y) \\ &= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\ &= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\ &= \prod_{i=1}^{n} p(x_i|y) \end{aligned}$$

Naive Bayes (NB) assumption:

- the x_i's are conditionally independent given y.

3 Naive Bayes

• 3.1 Example1: Sparm E-mail Classification (text classification)

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

- Our model is parameterized by $\phi_{i|y=1} = p(x_i = 1|y = 1)$, $\phi_{i|y=0} = p(x_i = 1|y = 0)$, and $\phi_y = p(y = 1)$. As usual, given a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, we can write down the joint likelihood of the data:

$$\mathcal{L}(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}).$$

- Maximizing this with respect to $\phi_y, \phi_{i|y=0}$ and $\phi_{i|y=1}$ gives the maximum likelihood estimates:

$$\begin{aligned}
\phi_{j|y=1} &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\
\phi_{j|y=0} &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
\phi_y &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}
\end{aligned}$$

3 Naive Bayes

• 3.2 Laplace smoothing

- Problem: Assuming that “nips” was the 35000th word in the dictionary, the Naive Bayes spam filter therefore had picked its maximum likelihood estimates of the parameters to be

$$\phi_{35000|y=1} = \frac{\sum_{i=1}^{m} 1\{x_{35000}^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} = 0$$

$$\phi_{35000|y=0} = \frac{\sum_{i=1}^{m} 1\{x_{35000}^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} = 0$$

- Laplace smoothing

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}{m} \rightarrow \phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} + 1}{m + k}.$$

3 Naive Bayes（朴素贝叶斯）

- **3.2 贝叶斯决策论（Bayesian decision theory）**

$$P(c \mid \mathbf{x})$$

  - “条件风险”（conditional risk）

$$R\left(c_i \mid \mathbf{x}\right)=\sum_{j=1}^{N} \lambda_{ij} P\left(c_j \mid \mathbf{x}\right)$$

  - 我们的任务是寻找一个判定准则 $h: X \mapsto Y$ 以最小化总体风险

$$R(h)=\mathbf{E}_x\left[R(h(\mathbf{x}) \mid \mathbf{x})\right]$$

3 Naive Bayes（朴素贝叶斯）

- **3.2 贝叶斯决策论（Bayesian decision theory）**
  - 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij} 可写为

$$\lambda_{i,j}\begin{cases} 0, & \text{if } i=j; \\ 1, & \text{otherwise}, \end{cases}$$

  - 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x})$$

  - 于是，最小化分类错误率的贝叶斯最有分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c \mid x)$$

  - 即对每个样本 $\mathbf{x}$，选择能使后验概率 $P(c \mid \mathbf{x})$ 最大的类别标记。

3 Naive Bayes（朴素贝叶斯）

- **3.2 贝叶斯决策论（Bayesian decision theory）**

“先验概率 + 新得到的证据 = 更正后的概率”

- 不受信息量多少的限制，将各种来源的结果，包括主观判断和有限的客观信息，综合到一起，得到最后的结论。

先验概率：样本空间中各类样本所占的比例，可通过各类样本出现的频率估计

类标记c相对于样本x的“类条件概率”，或称“似然”。

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$$

后验概率：结合先验概率与条件概率，得到的属于类别c的概率估计

“证据”因子，与类标记无关

3 Naive Bayes

- **Example2：给定一幅图像，判定其拍摄季节是秋天还是春天？**

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

3 Naive Bayes

- **Example3: 七夕要不要送女/男朋友礼物？**

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

友情提示：还有三周

3 Naive Bayes

- **Example3: 七夕要不要送女/男朋友礼物?**

3 Naive Bayes

- **Example4: 海洋搜救**

- John Craven
  - 数学家
  - 美国海军特别计划部首席科学家
  - 采用贝叶斯方法，搜寻美国海军在汪洋大海里搜索丢失的氢弹（1966年）、失踪的核潜艇（1968年）

3 Naive Bayes

- **天蝎号核潜艇**
  - 事后调查报告：
  - 罪魁祸首是这艘潜艇上的一枚奇怪的鱼雷
  - 发射出去后竟然敌我不分，扭头射向自己，让潜艇中弹爆炸。

Photo # NH 70305 USS Scorpion comes alongside USS Tallahatchie County, April 1968

3 Naive Bayes

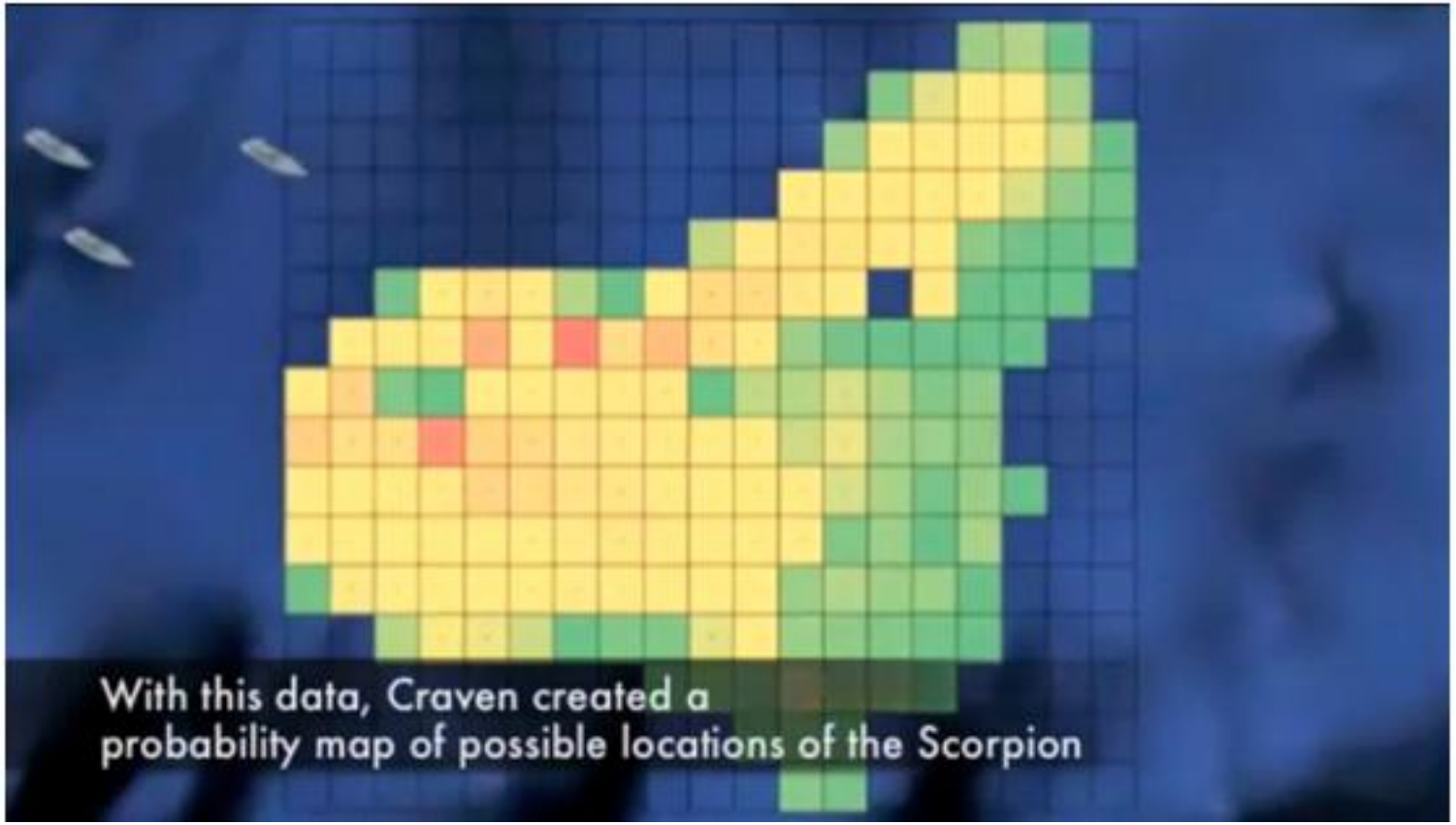
- **搜救困难**
  - 失事时潜艇航行的速度快慢，方向，爆炸冲击力的大小方向，爆炸时潜艇方向舵的指向都是未知量
  - 即使知道潜艇在哪里爆炸，也很难确定潜艇残骸最后被海水冲到哪里。
  - Craven初略地估计一下，在半径20英里的圆圈内的海底，天蝎潜艇都有可能躺在那里，要在这么大的范围内、这么深的海底找到潜艇几乎成了不可能完成的任务。

Photo # NH 70305 USS Scorpion comes alongside USS Tallahatchie County, April 1968

3 Naive Bayes

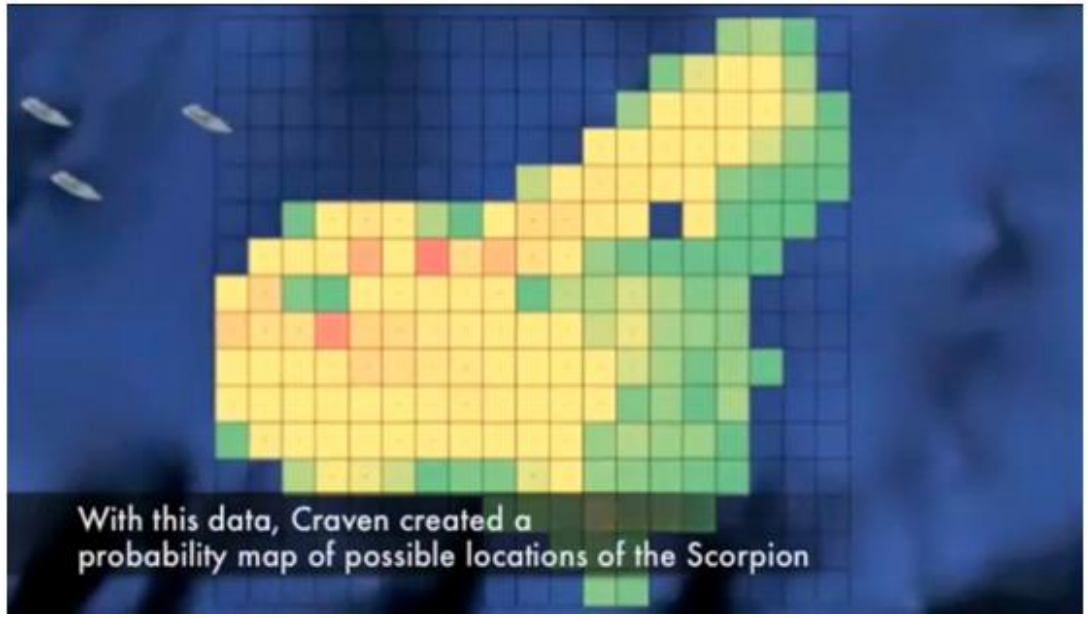
- **主观猜测**
  - Craven咨询了数学家、潜艇专家、海事搜救各个领域的专家
  - 编写了各种可能的“剧本”
  - 按照自己的知识和经验对于情况会向哪一个方向发展进行猜测。

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$



3 Naive Bayes

- **概率校正**
  - 每个小格子有两个概率值p和q，
    - p：潜艇躺在这个格子里的概率
    - q：如果潜艇在这个格子里，它被搜索到的概率。



  - 按照经验，第二个概率值主要跟海域的水深有关，在深海区域搜索时失事潜艇“漏网”的可能性会更大。
  - 如果一个格子被搜索后，没有发现潜艇的踪迹，按照贝叶斯原理更新后，这个格子潜艇存在的概率就会降低：

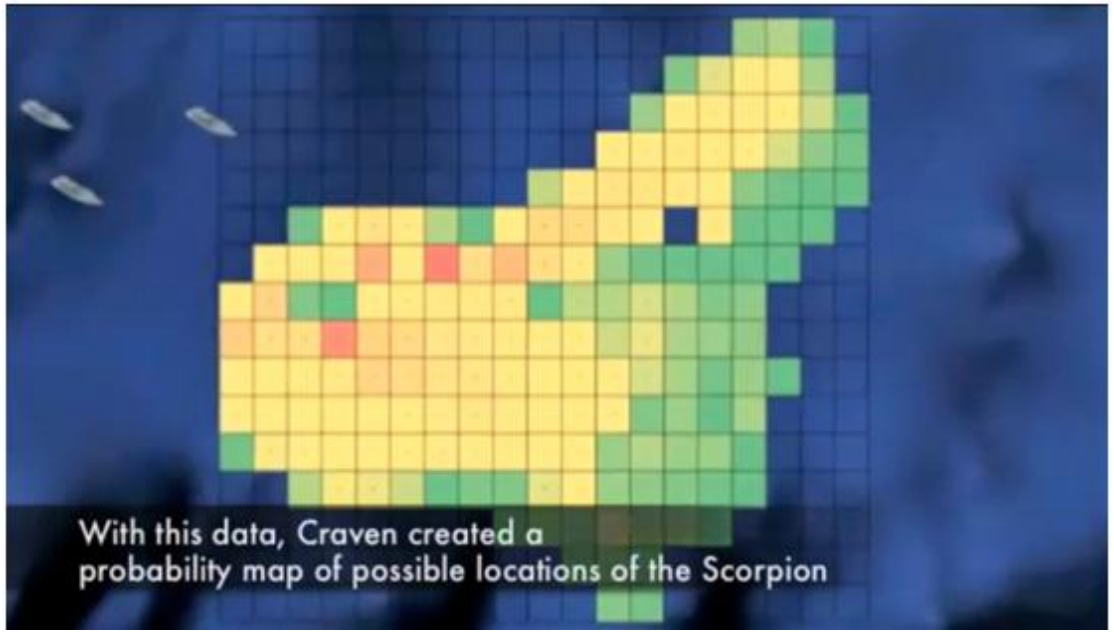
3 Naive Bayes

- **概率校正**
  - 如果一个格子被搜索后，没有发现潜艇的踪迹，按照贝叶斯原理更新后，这个格子潜艇存在的概率就会降低：

$$p' = \frac{p(1-q)}{(1-p)+p(1-q)} = p\frac{1-q}{1-pq} < p.$$

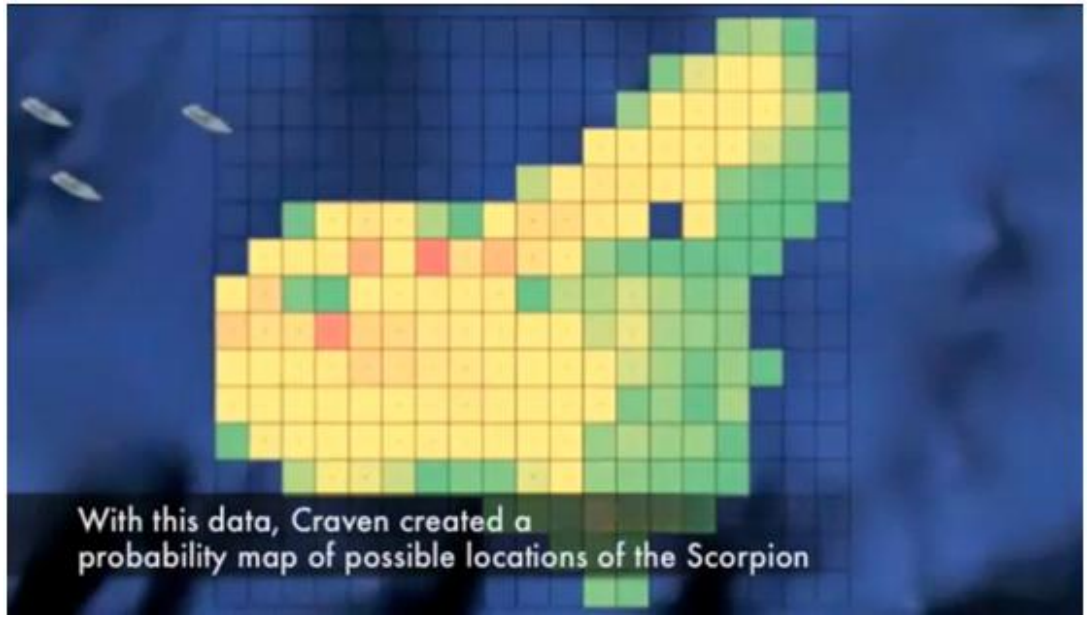
  - 其他各个格子的潜艇存在的概率值就会上升：

$$r' = r\frac{1}{1-pq} > r.$$



3 Naive Bayes

- **概率校正**
  - 每次寻找时会挑选整个区域内潜艇存在概率值最高的一个格子进行搜索，
  - 如果没有发现，概率分布图会被“洗牌”一次，
  - 搜寻船只就会驶向新的“最可疑格子”进行搜索，
  - 这样一直下去，直到找到天蝎为止。



最初的时候，海军人员凭经验估计潜艇是在爆炸点的东侧海底，但是几个月的搜索后一无所获。后来不听从了Craven的建议，按照概率图，失事后的潜艇应该在爆炸点的西侧。经过几次搜索，在爆炸点西南方的海底找到了。

目录

Thank You !

Q & A