Classification: Three Different Methods

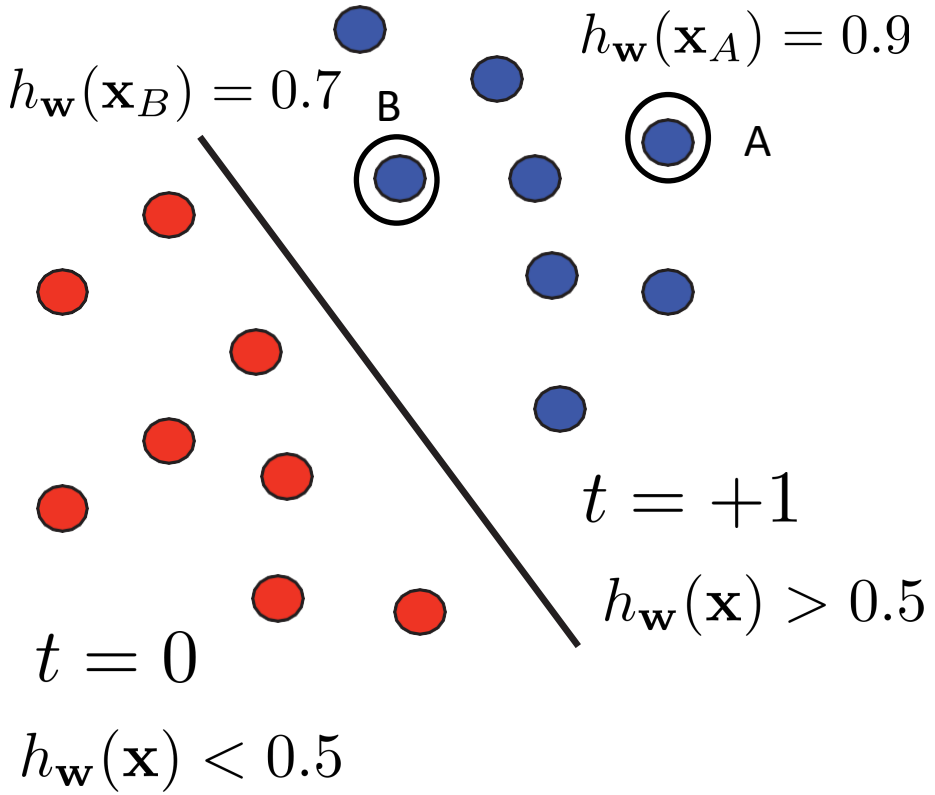
- **Discriminant models**
  - Given training data, assign each data x to one class C_k via a discriminant function
  - *Do not consider distribution* of the training data

- **Probabilistic discriminant models**
  - Given training data, model the posterior class distribution $p(C_k|x)$
  - Use the distribution $p(C_k|x)$ to perform classification for testing data

- **Probabilistic generative models**
  - Given training data, model the joint (data, class) distribution $p(x,C_k)$
  - Find class-conditional distribution $p(x|C_k)$ and class prior distribution $p(C_k)$
  - Then use Bayes rule to compute $p(C_k|x) \sim p(x|C_k)\,p(C_k)$

Probabilistic Generative Models: Gaussian Discriminant Analysis & Naïve Bayes Classifier

Recap: Logistic Regression

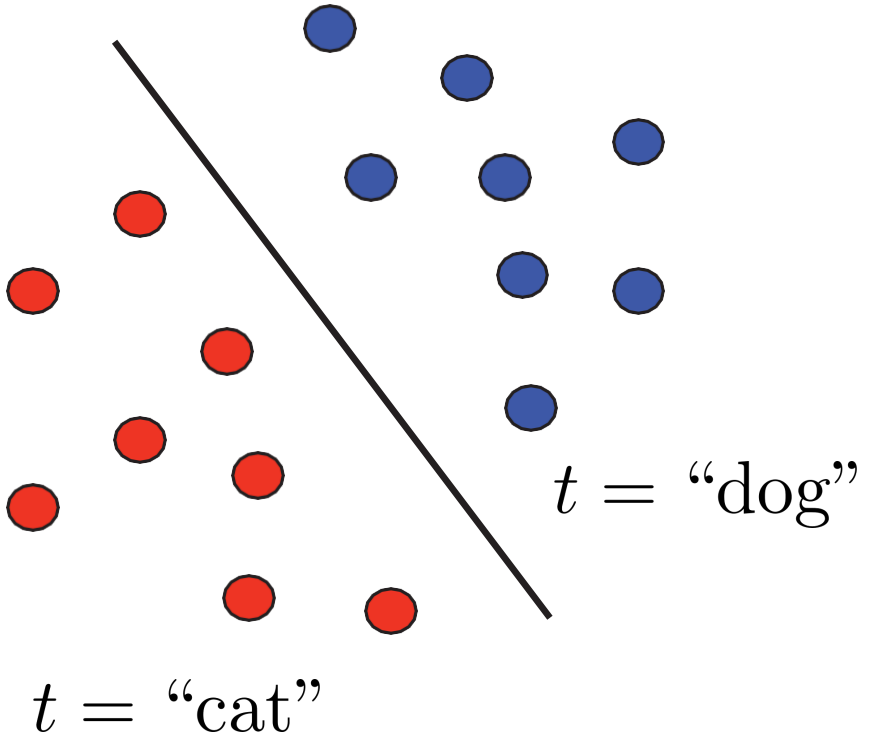


Model conditional/posterior distribution of t given x

$$p(t|\mathbf{x};\mathbf{w}) = ?$$

$$p(t=1|\mathbf{x};\mathbf{w})$$
$$= \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$$
$$= h_{\mathbf{w}}(\mathbf{x})$$

Classification: Different Perspectives



Consider a classification problem: distinguish **dogs** (t = 1) vs **cats** (t = 0) based on features of an animal

Logistic regression: find a decision boundary to separate **cats** and **dogs**, via modeling **cat/dog label posterior distribution** given animal features

$$p(t = \text{“dog”}|\mathbf{x}) \textbf{ vs } p(t = \text{“cat”}|\mathbf{x})$$

A different perspective

- Look at dogs to build a model of what dogs look like
- Look at cats to build another model of what cats look like
- To classify a new animal, we match it to the dog/cat model, and see whether it looks more like the dog or cat

$$p(\mathbf{x}|t = \text{“dog”}) \textbf{ vs } p(\mathbf{x}|t = \text{“cat”})$$

Probabilistic Generative vs. Discriminative Models

Discriminative models: $p(t|\mathbf{x})$ Label posterior probability

Learn mappings directly from the space of inputs to the space of labels

Generative models: $p(\mathbf{x}|t)$ Likelihood $p(t)$ Label prior

Data generation process: model distribution of likelihood & label prior

Bayes rule to derive label t's posterior distribution given x

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t) \cdot p(t)}{p(\mathbf{x})}$$

$p(\mathbf{x})$ Normalization constant

Prediction $\arg\max_t p(t|\mathbf{x}) = \arg\max_t p(\mathbf{x}|t)p(t)$

Preliminaries

Bayes Rule

Joint probability distribution

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

P(A|B) (P(B|A)): **Conditional probability** of event A (B) given event B (A) happens

P(A) (P(B)): **Prior probability** of observing event A (B)

Bayes Rule $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of total probability $$P(B) = \sum_a P(A = a, B) = \sum_a P(B|A = a)P(A = a)$$

Multivariate Gaussian/Normal Distribution

$$p(\mathbf{x} \in \mathbb{R}^d; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$$

Mean vector: $\mu \in \mathbb{R}^d$ $\quad \mathbb{E}(\mathbf{x}) = \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}; \mu, \Sigma) d\mathbf{x} = \mu$

Covariance matrix: $\Sigma \in \mathbb{R}^{d \times d}$ $\quad \mathrm{Cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \Sigma$

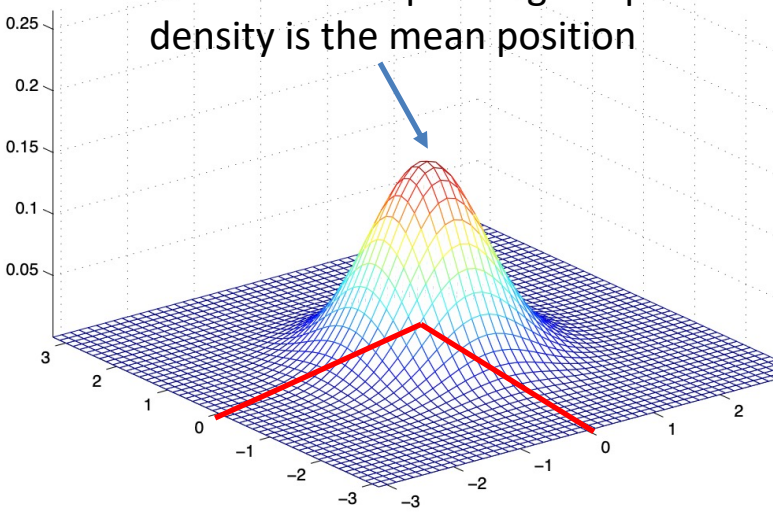
Determinant: $|\Sigma|$

Symmetric: $\Sigma^T = \Sigma$

Positive semi-definite: $\mathbf{v}^T \Sigma \mathbf{v} \geq 0, \forall \mathbf{v} \in \mathbb{R}^d$

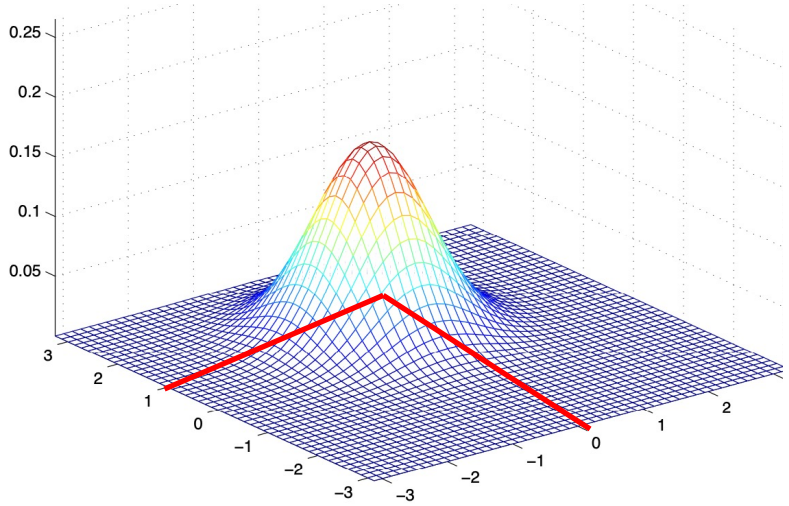
Density of Gaussian Distributions

Position corresponding the peak density is the mean position

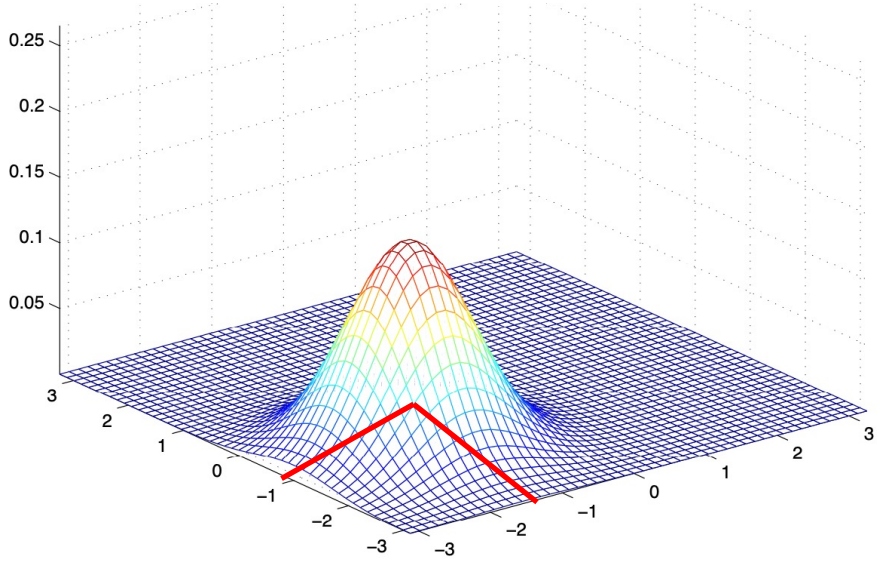


$$\mathcal{N}([0;0], \Sigma = \mathbf{I})$$

Standard Gaussian distribution



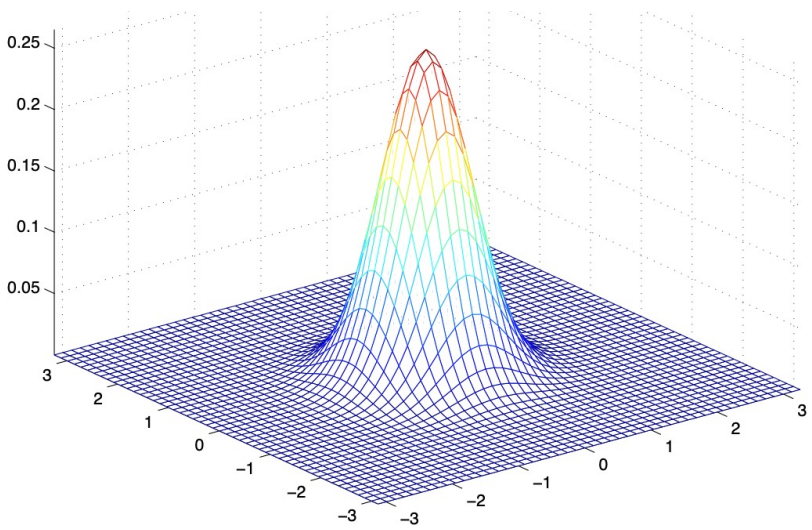
$$\mathcal{N}([1;0]; \Sigma = \mathbf{I})$$



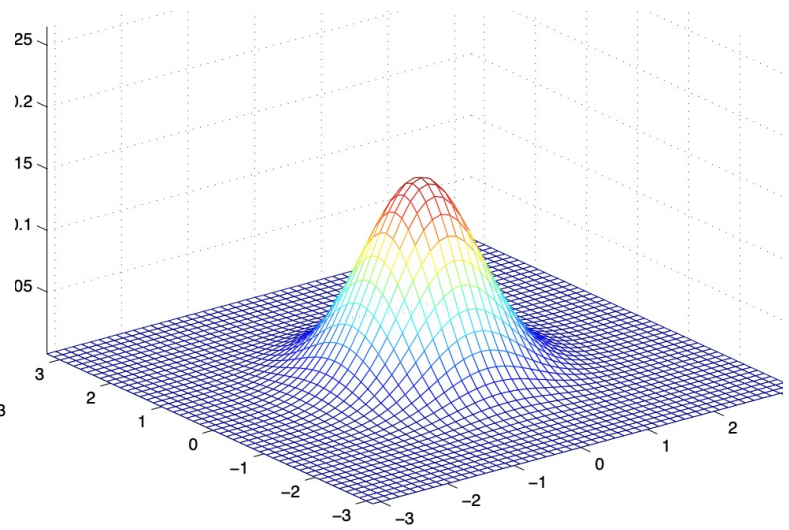
$$\mathcal{N}([-1;-1.5]; \Sigma = \mathbf{I})$$

Varying mean μ, move the peak positions of the density, but the density shapes are the same

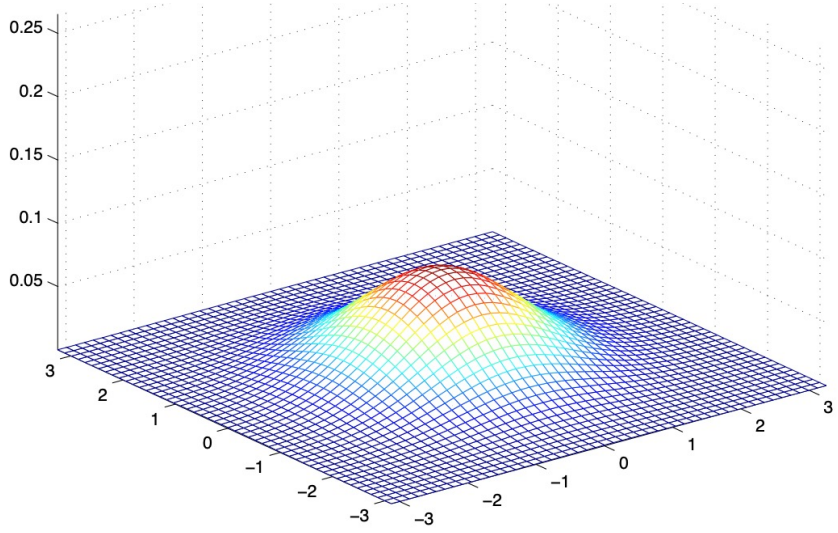
Density of Gaussian Distributions



$$\mathcal{N}([0;0], \Sigma = 0.6\mathbf{I})$$



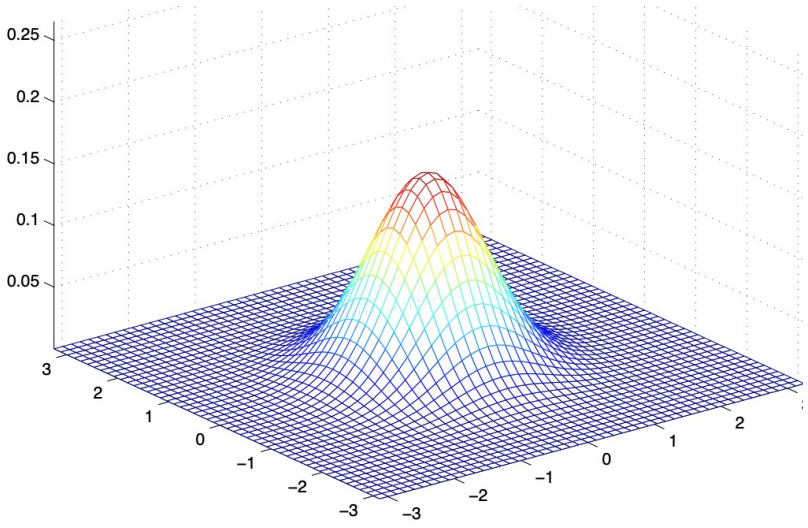
$$\mathcal{N}([0;0], \Sigma = \mathbf{I})$$



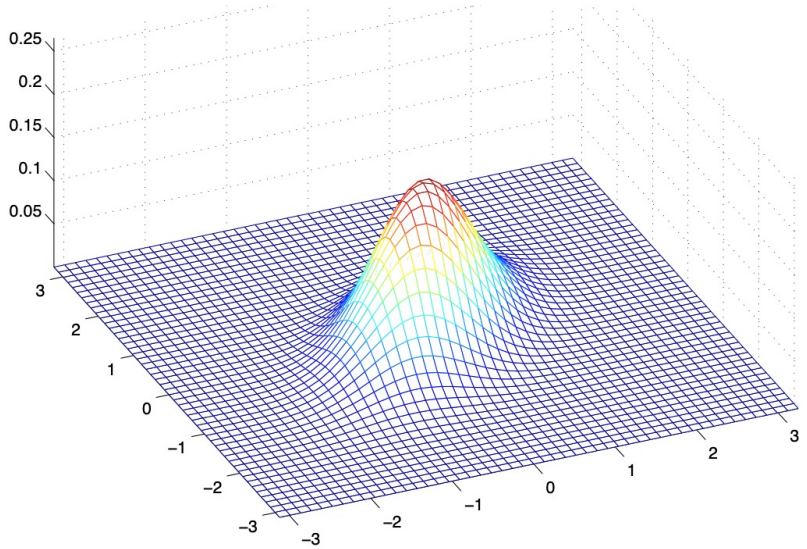
$$\mathcal{N}([0;0]\Sigma = 2\mathbf{I})$$

Σ ***larger*** (*smaller*), Gaussian becomes more spread-out (compressed)

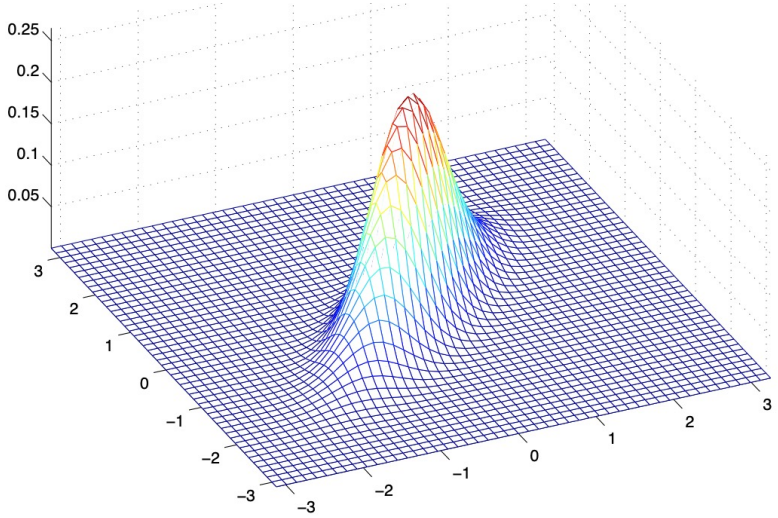
Density of Gaussian Distributions



$$\mathcal{N}([0;0], \Sigma = \mathbf{I})$$



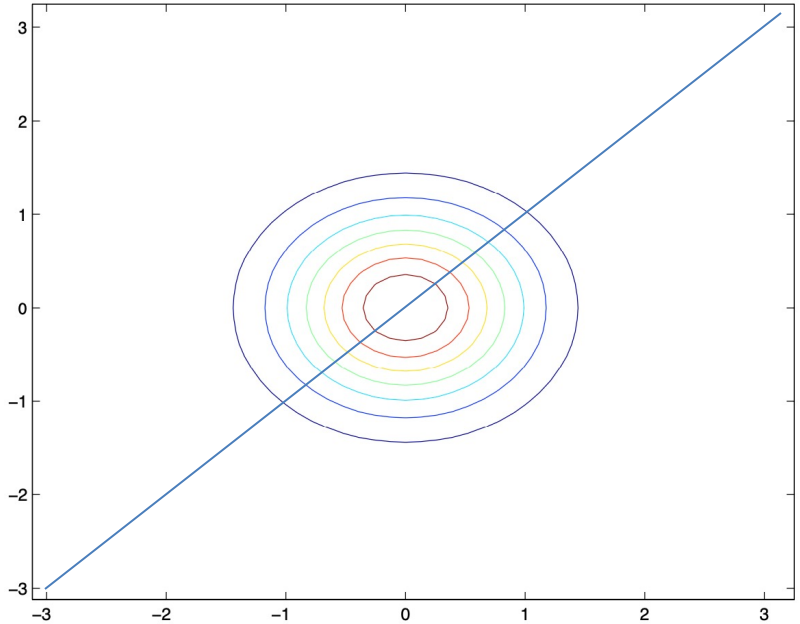
$$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$$



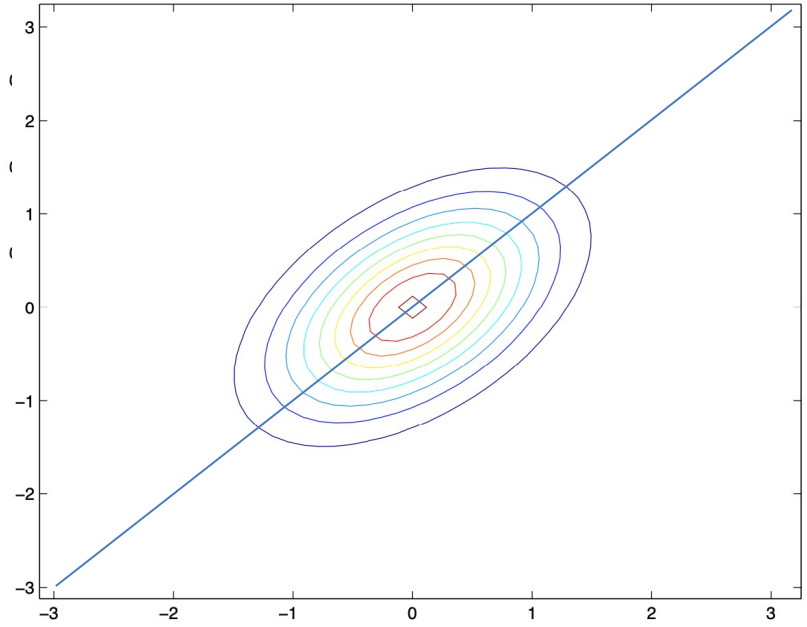
$$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix})$$

Increasing the off-diagonal entry in Σ, the density becomes more "compressed" towards the $x_2 = x_1$ line

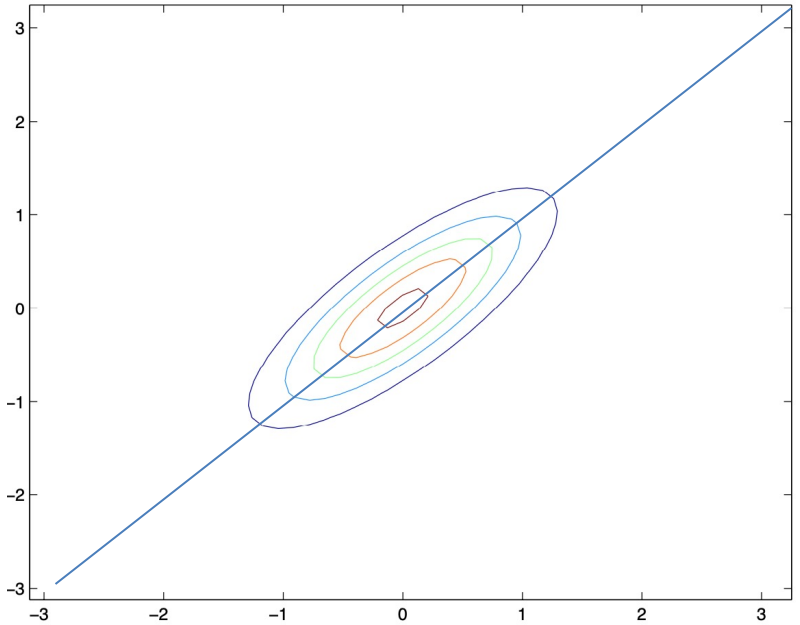
Contour of Density



$\mathcal{N}([0;0], \Sigma = \mathbf{I})$



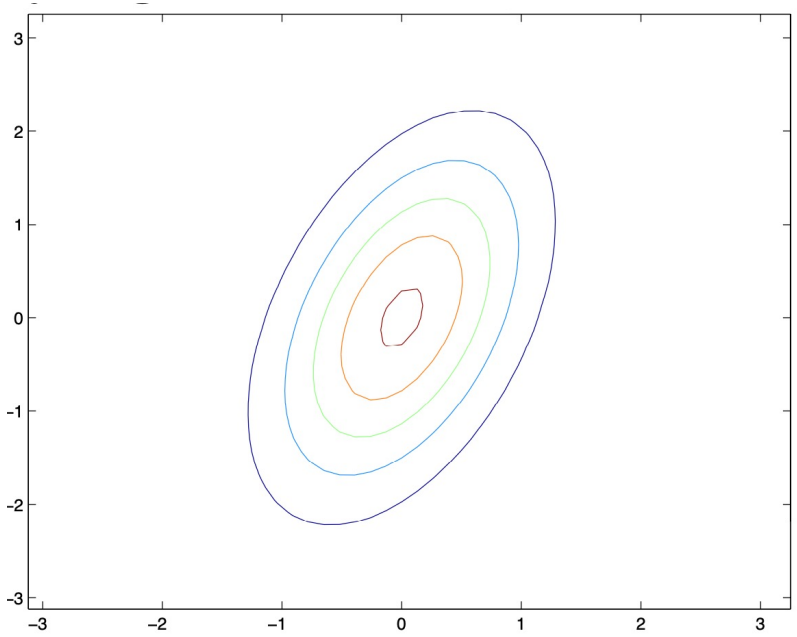
$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$

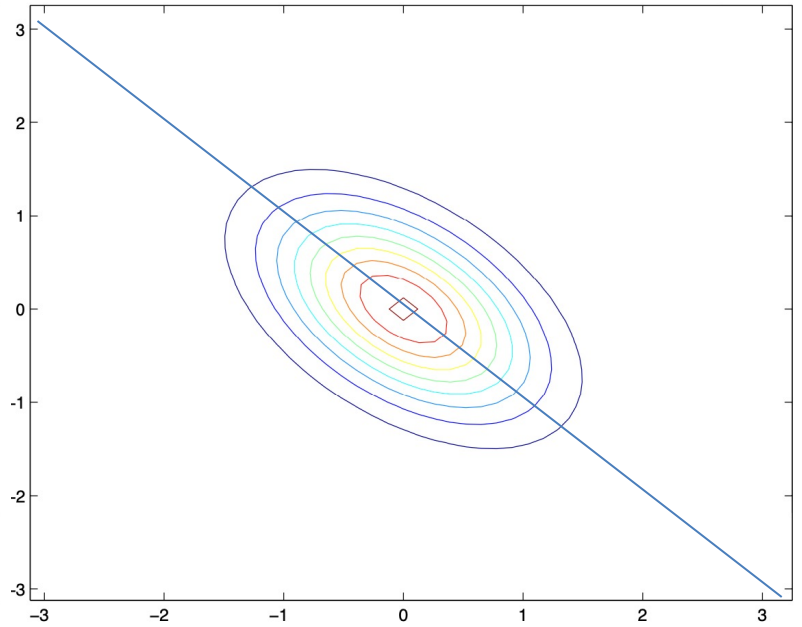


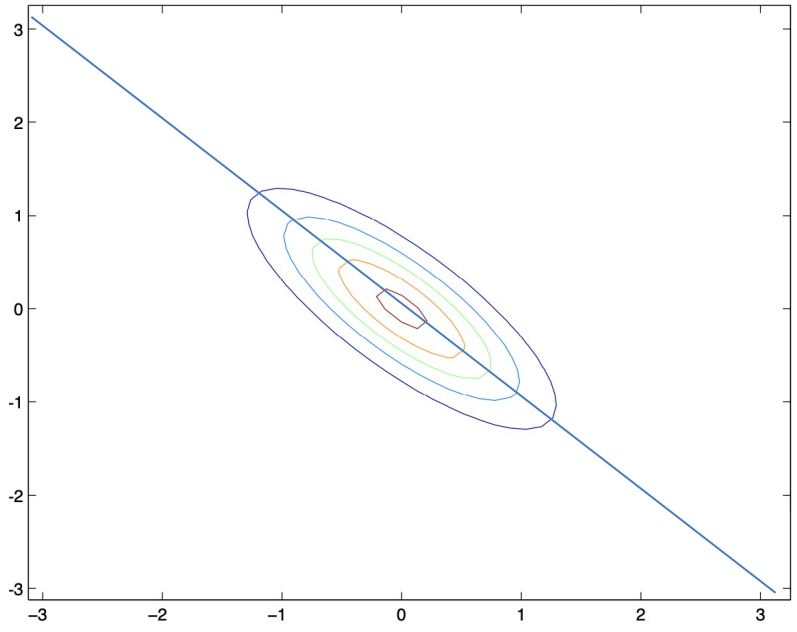
$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix})$

Increasing the off-diagonal entry in Σ, the density becomes more "compressed" towards the $x_2 = x_1$ line

Contour of Density







$$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix})$$

$$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

$$\mathcal{N}([0;0], \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix})$$

Decreasing the off-diagonal entry in Σ, the density becomes more “compressed” towards the $x_2 = -x_1$ line

Generally, the contours form ellipses by varying Σ parameters

Gaussian Discriminant Analysis (GDA)

Probabilistic Generative Models

Generative models: $p(\mathbf{x}|t)$ Likelihood $p(t)$ Label prior

Data generation process: model distribution of likelihood & label prior

Bayes rule to derive label t's posterior distribution given x

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t) \cdot p(t)}{p(\mathbf{x})}$$

Prediction $\arg\max_t p(t|\mathbf{x}) = \arg\max_t p(\mathbf{x}|t)p(t)$

Gaussian Discriminant Analysis: Continuous Random Variables

Binary classification t ∈ {0,1}

Likelihood

$$\mathbf{x}|t=0 \sim \mathcal{N}(\mu_0, \Sigma) \qquad p(\mathbf{x}|t=0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0)\right)$$

$$\mathbf{x}|t=1 \sim \mathcal{N}(\mu_1, \Sigma) \qquad p(\mathbf{x}|t=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right)$$

Note that there are two different mean vectors but just one covariance matrix

Label Prior

$$t \sim \text{Ber}(\phi) \qquad p(t) = \phi^t(1-\phi)^{1-t}$$

Parameters: $\phi, \mu_0, \mu_1, \Sigma$

(Log-) Likelihood of Data

Given N independent data samples $\{(x_n, t_n)\}_{n=1}^N$, the likelihood is defined as

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \log \prod_{n=1}^{N} p(\mathbf{x}_n|t_n) \cdot p(t_n)$$

$$= \sum_{n=1}^{N} \log p(\mathbf{x}_n|t_n) + \sum_{n=1}^{N} \log p(t_n)$$

$$= \sum_{n:t_n=0} \log \Big[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0)\Big)\Big]$$

$$+ \sum_{n:t_n=1} \log \Big[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\Big)\Big]$$

$$+ \sum_{n=1}^{N} \phi^{t_n}(1-\phi)^{1-t_n}$$

Maximum Likelihood Estimation

$$\frac{\partial \ell}{\partial \phi} = 0 \qquad \phi = \frac{1}{N}\sum_{n=1}^{N} 1_{[t_n=1]}$$

Fraction of data points belonging to class *1*

$$\frac{\partial \ell}{\partial \mu_i} = 0 \qquad \mu_i = \frac{\sum_{n=1}^{N} 1_{[t_n=i]} \cdot \mathbf{x}_n}{\sum_{n=1}^{N} 1_{[t_n=i]}}$$

Average features of data points belonging to class *i*

$$\frac{\partial \ell}{\partial \Sigma} = 0 \qquad \Sigma = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n - \mu_{t_n}\right)\left(\mathbf{x}_n - \mu_{t_n}\right)^T$$

Covariance matrix

Obtain: $p(\mathbf{x}|t = 0/1; \mu_{0/1}, \Sigma) \quad p(t; \phi)$

Prediction: $\arg\max_t p(t|\mathbf{x}_{te}) = \arg\max_t p(\mathbf{x}_{te}|t)p(t)$

$$p(\mathbf{x}_{te}|t = 0)p(t = 0) \textbf{ vs } p(\mathbf{x}_{te}|t = 1)p(t = 1)$$

Training: GDA with MLE

Input: Giving a set of N data points $\{(x_n, t_n)\}_{n=1}^{N}$ from two classes

Step 0: Model the likelihood of each class of data points as a Gaussian distribution

Step 1: Estimate the mean and covariance of two Gaussian distributions

$$\mu_i = \frac{\sum_{n=1}^{N} 1_{[t_n=i]} \cdot \mathbf{x}_n}{\sum_{n=1}^{N} 1_{[t_n=i]}},\ i \in \{0, 1\}$$

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - \mu_{t_n}\right)\left(\mathbf{x}_n - \mu_{t_n}\right)^T$$

Step 2: Estimate the prior p(t=k) as the fraction of training samples with label t=k

$$p(t = k) = \phi_k = \frac{1}{N} \sum_{n=1}^{N} 1_{[t_n=k]}$$

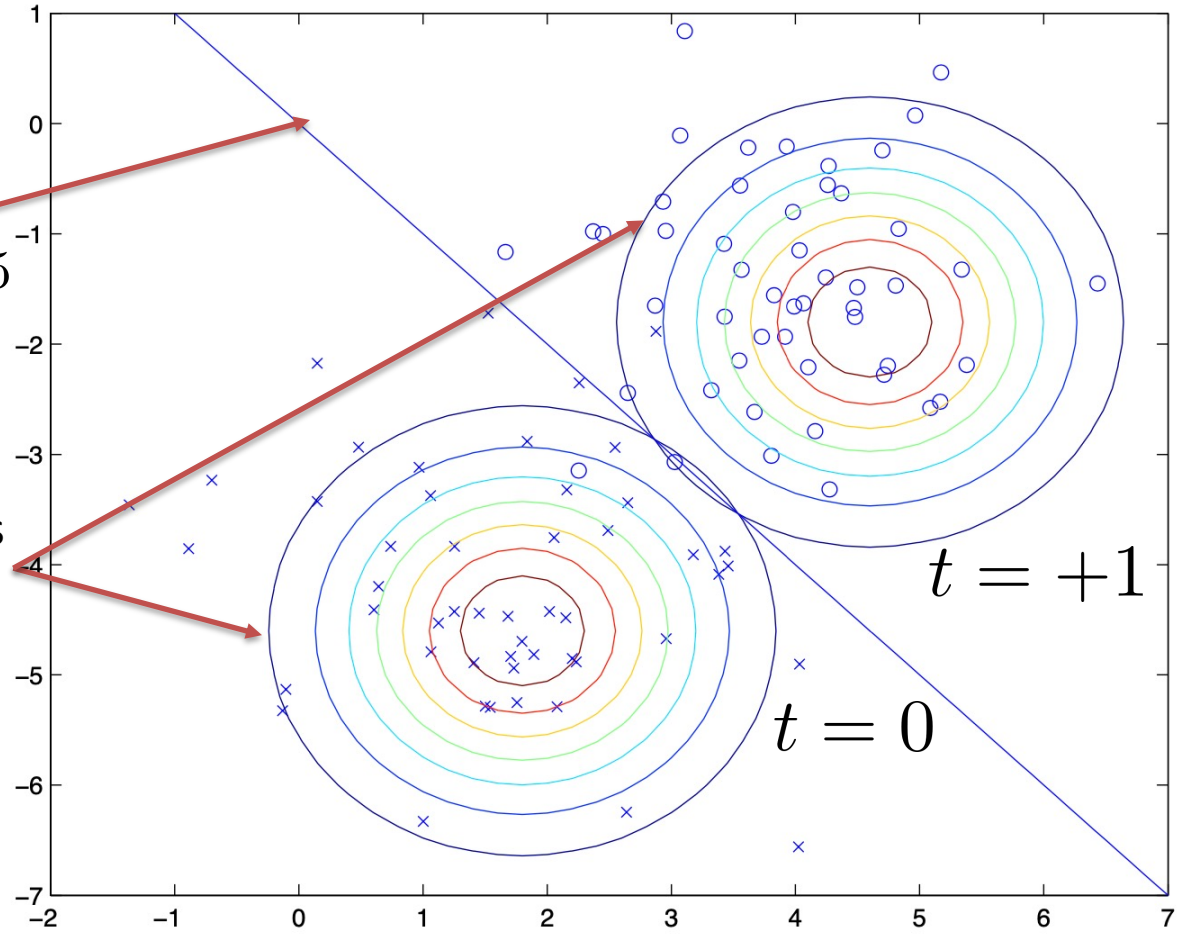
An Example

Decision boundary

$$p(t=1|\mathbf{x}) = p(t=0|\mathbf{x}) = 0.5$$

Contours of two gaussian dists fit to the data for two classes

Same shape and orientation due to the same cov. matrix



$$p(\mathbf{x}_{te}|t=0)p(t=0) \textbf{ vs } p(\mathbf{x}_{te}|t=1)p(t=1)$$

Naive Bayes Classifier (NBC): Discrete Random Variables

Motivating Example: Spam Filtering

Sir / Madam,

We invite you to submit your manuscript(s) for publication. The journals include research papers, review articles, technical projects and short communications containing new insight into any aspect of the covered scope of the journal. Our objective is to inform authors of the decision on their manuscript(s) within weeks of submission. After acceptance, the paper will be published in the current issue immediately.
Keywords: English, Literature, Science, Economics, Engineering, Management, Agriculture, Horticulture, Environment .....

International Journal of Advanced Engineering Research and Science (IJAERS) ISSN: 2456-1908(O) | 2349-6495 (P)

DOI (CrossRef): 10.22161/ijaers

Thomson Reuters ResearcherID: P-3738-2015

Impact Factor: 4.192, SJIF: 4.072, IBI: 3.2, PIF: 2.465, ISRA-JIF: 1.317,

Website: http://www.ijaers.com

Kindly submit research articles to **http://ijaers.com/submit-paper/** or mail us at **editor.ijaers@gmail.com**

International Journal of English, Literature and Science (IJELS) ISSN: 2456-7620

New Email

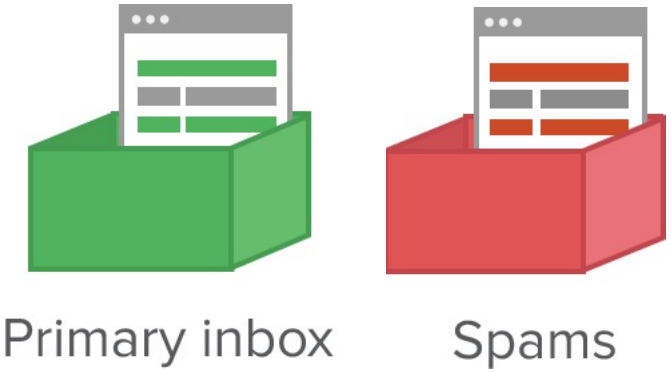
Classifier

Predict

Is Spam

Train

Binary classification



Primary inbox

Spams

Emails with Labels

Motivating Example: Spam Filtering

Represent an email via a feature vector whose length is equal to the **number of words** in the dictionary

If an email contains the i-th word of the dictionary, then we will set xi = 1; otherwise, we let xi = 0

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix} \in \{0,1\}^{50,000}$$

Now we build a generative model to learn p(**x**|t)

How many parameters to learn p(**x**|t) ?

$$2 \cdot 2^{50,000}$$

Computationally infeasible

Naïve Bayes Assumption

Naïve Bayes Assumption: All features are ***independent*** given the class label t

$$p(x_1, x_2, x_{50000}|t) = \prod_{i=1}^{50000} p(x_i|t) \qquad \#\ \text{paras.:}\ 2 \cdot 50,000$$

Strong assumption but works well in practice!

E.g., , if t = 1 (spam), “buy” is word 100 and “price” is word 800.

If I tell you t = 1, knowing the value of x100 (whether “buy” or not) has no effect on your beliefs about the value of x800 (whether “price” appears)

$$p(x_{100}, x_{800}|t) = p(x_{100}|t) \cdot p(x_{800}|t, x_{100}) = p(x_{100}|t) \cdot p(x_{800}|t)$$

Note: This is different from that x100 and x800 are independent

$$p(x_{100}) = p(x_{100}|x_{800})$$

Naïve Bayes Model: Binary Random Variables

Binary random variable per feature: $x_i \in \{0,1\}\ \forall i$

$$x_i|t = k \sim \text{Ber}(\phi_{i|t=k}) \qquad p(x_i|t = k; \phi_{i|t=k}) = \phi_{i|t=k}^{x_i}(1 - \phi_{i|t=k})^{1-x_i}$$

Binary/multiclass classification t ∈ {0,1}/{1,…K}

Binary classification

$$t \sim \text{Ber}(\phi) \qquad p(t; \phi) = \phi^t(1 - \phi)^{1-t}$$

Multiclass classification

$$t \sim \text{Multi}(\{\phi_k\}_{k=1}^K) \qquad p(t; \{\phi_k\}) = \phi_1^{1_{y=1}} \phi_2^{1_{y=2}} \cdot \phi_K^{1_{y=K}}$$

$$\sum_k \phi_k = 1$$

Parameters: $\{\phi_{i|t=k}\}, \phi(\phi_k)$

(Log-) Likelihood of Data: Binary Classification

Given N independent data samples $\{(x_n, t_n)\}_{n=1}^N$, the likelihood is defined as

$$\max \ell(\phi, \{\phi_{i|t=k}\}) = \log \prod_{n=1}^{N} p(\mathbf{x}_n, t_n) = \sum_{n=1}^{N} \log p(\mathbf{x}_n, t_n)$$

$$= \sum_{n=1}^{N} \log p(\mathbf{x}_n | t_n) + \sum_{n=1}^{N} \log p(t_n)$$

Naïve Bayes assumption

$$= \sum_{n=1}^{N} \sum_{i=1}^{d} \log p(x_{n,i} | t_n) + \sum_{n=1}^{N} \log p(t_n)$$

$$\text{subject to } \sum_{t \in \{0,1\}} p(t) = 1, \quad p(t) \geq 0$$

$$\sum_{x_i \in \{0,1\}} p(x_i|t) = 1,\ p(x_i|t) \geq 0, \forall i \in \{1, 2, \cdots, d\},\ t \in \{0, 1\}$$

Lagrangian

$$\max \ell(\phi, \{\phi_{i|t}\}; \alpha, \beta) = \sum_{n=1}^{N} \sum_{i=1}^{d} \log p(x_{n,i}|t_n) + \sum_{n=1}^{N} \log p(t_n) + \alpha \Big(1 - \sum_{t \in \{0,1\}} p(t)\Big) + \sum_{t \in \{0,1\}} \sum_{i=1}^{n} \beta_i \Big(1 - \sum_{x_i \in \{0,1\}} p(x_i|t)\Big)\Big)$$

Maximum Likelihood Estimation (MLE)

$$\frac{\partial \ell}{\partial \phi} = 0 \qquad \phi = \frac{1}{N}\sum_{n=1}^{N} 1_{[t_n=1]}$$

Fraction of data points belonging to class *1*

$$\frac{\partial \ell}{\partial \phi_{i|t=0}} = 0 \qquad \phi_{i|t=0} = \frac{\sum_{n=1}^{N} 1_{[t_n=0 \& x_{n,i}=1]}}{\sum_{n=1}^{N} 1_{[t_n=0]}}$$

$$\frac{\partial \ell}{\partial \phi_{i|t=1}} = 0 \qquad \phi_{i|t=1} = \frac{\sum_{n=1}^{N} 1_{[t_n=1 \& x_{n,i}=1]}}{\sum_{n=1}^{N} 1_{[t_n=1]}}$$

Fraction of data points belonging to class *0/1,* AND their i-th feature is 1

Obtain: $p(\mathbf{x}|t) = \prod_{i=1}^{d} p(x_i|t;\phi_{i|t}) \qquad p(t;\phi)$

Prediction: $\arg\max_t p(t|\mathbf{x}_{te}) = \arg\max_t p(\mathbf{x}_{te}|t)p(t)$

$$p(\mathbf{x}_{te}|t=0)p(t=0) \textbf{ vs } p(\mathbf{x}_{te}|t=1)p(t=1)$$

Maximum Likelihood Estimation (Details)

$$\max \ell(\phi, \{\phi_{i|t}\}; \alpha, \beta) = \sum_{n=1}^{N}\sum_{i=1}^{d} \log p(x_{n,i}|t_n) + \sum_{n=1}^{N} \log p(t_n) + \alpha\Big(1 - \sum_{t\in\{0,1\}} p(t)\Big) + \sum_{t\in\{0,1\}}\sum_{i=1}^{n} \beta_i\Big(1 - \sum_{x_i\in\{0,1\}} p(x_i|t)\Big)$$

$$\frac{\partial \ell}{\partial p(t)} = \sum_{n:t_n=t} \frac{\partial \log p(t)}{\partial p(t)} - \alpha = \sum_{n:t_n=t} \frac{1}{p(t)} - \alpha = \frac{\sum_n 1_{[t_n=t]}}{p(t)} - \alpha = 0$$

Fraction of data points belonging to class *1*

$$\Rightarrow \quad p(t) = \frac{\sum_{n=1}^{N} 1_{[t_n=t]}}{\alpha}$$

$$\sum_t p(t) = 1 \quad \Rightarrow \quad \alpha = N \quad \Rightarrow \quad p(t) = \frac{1}{N}\sum_{n=1}^{N} 1_{[t_n=t]} \quad \Rightarrow \quad \phi = p(t=1) = \frac{1}{N}\sum_{n=1}^{N} 1_{[t_n=1]}$$

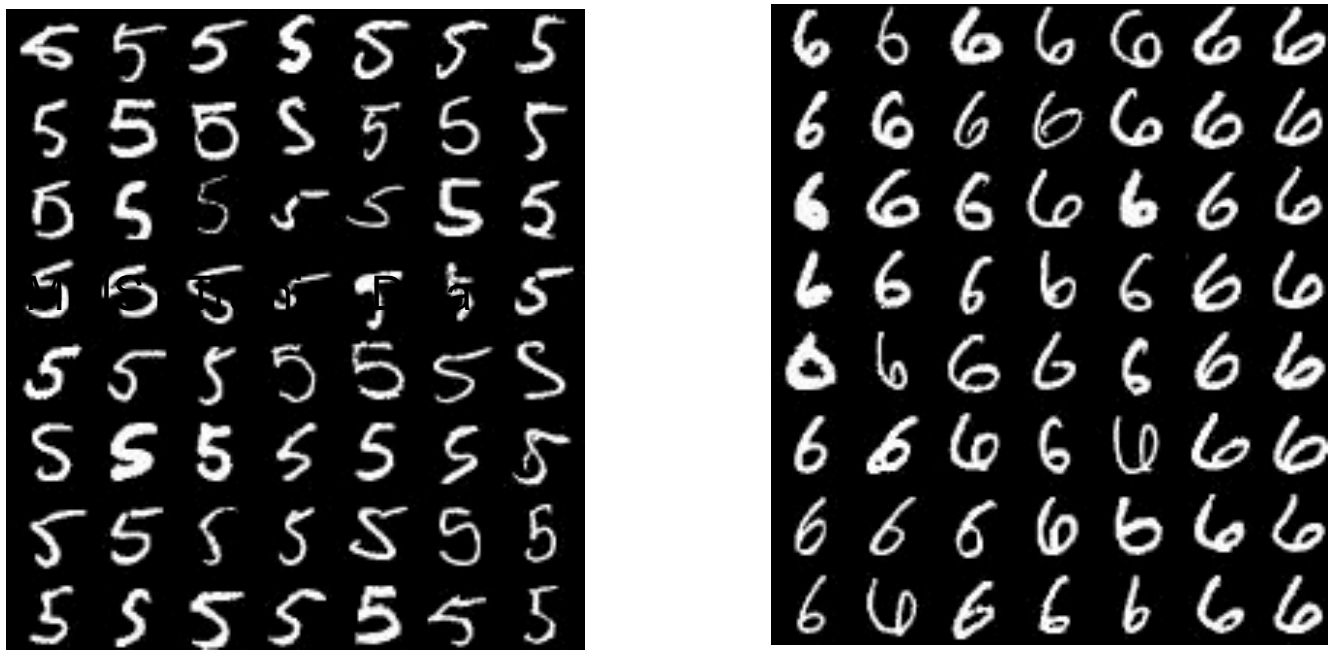
$$\frac{\partial \ell}{\partial p(x_i|t)} = \sum_{n:x_{n,i}=x_i, t_n=t} \frac{\partial \log p(x_i|t)}{\partial p(x_i|t)} - \beta_i = \frac{\sum_n 1_{[x_{n,i}=x_i \& t_n=t]}}{p(x_i|t)} - \beta_i = 0$$

Fraction of data points belonging to a class *k*, AND their i-th feature is 1

$$\Rightarrow \quad p(x_i|t) = \frac{\sum_n 1_{[t_n=t \& x_{n,i}=x_i]}}{\beta_i}$$

$$\sum_{x_i} p(x_i|t) = 1, \forall t \quad \Rightarrow \quad \beta_i = \sum_n 1_{[t_n=t]} \quad \Rightarrow \quad \phi_{x_i|t} = p(x_i|t) = \frac{\sum_{n=1}^{N} 1_{[t_n=t \& x_{n,i}=x_i]}}{\sum_{n=1}^{N} 1_{[t_n=t]}}$$

Example: Binary Digit Recognition

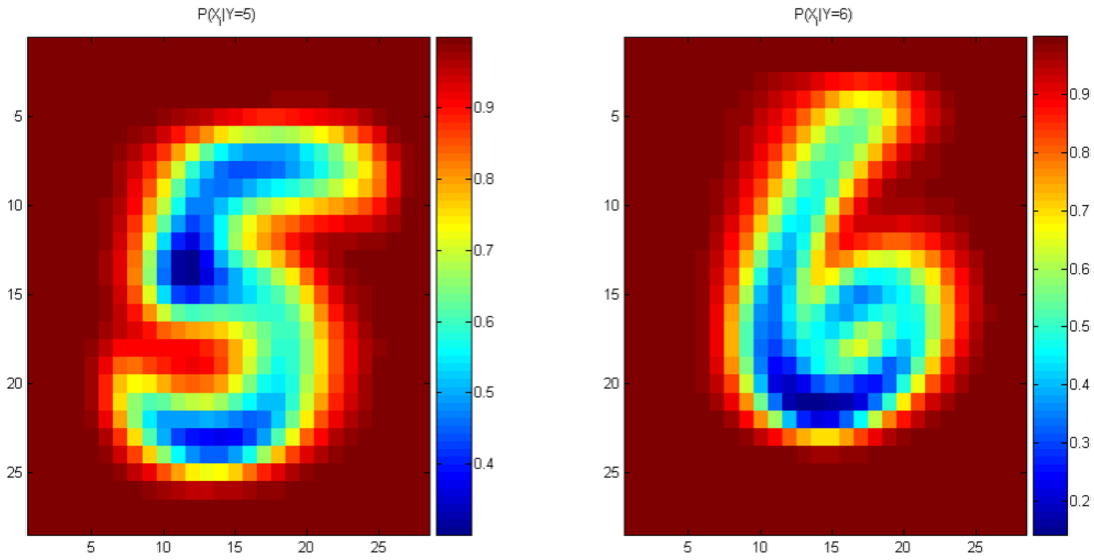


MNIST Training Data

Training: Naïve Bayes with MLE

Step 1: Estimate likelihood $P(X_i=0/1|t=k)$ as the fraction of training samples with $t=k$ for which $X_i=0/1$

$$p(x_i|t=k) = \phi_{x_i|t=k} = \frac{\sum_{n=1}^{N} 1_{[t_n=k \& x_{n,i}=x_i]}}{\sum_{n=1}^{N} 1_{[t_n=k]}}$$



Step 2: Estimate prior $P(t=k)$ as the fraction of training samples with label $t=k$

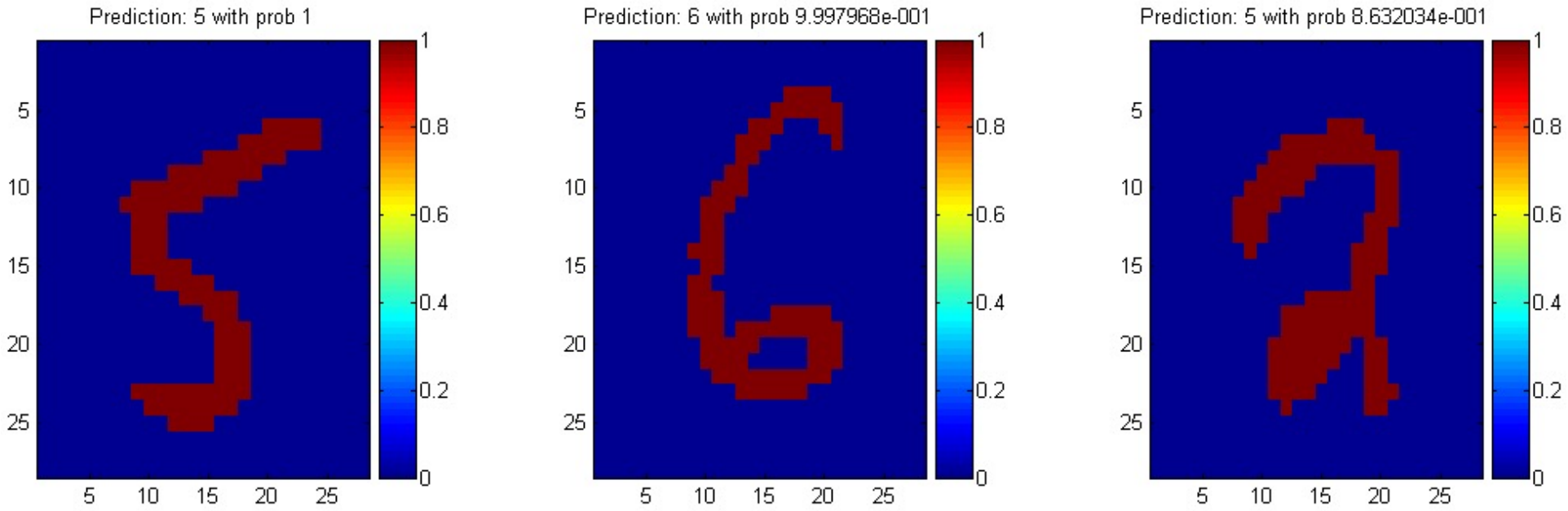
$$p(t=k) = \phi_k = \frac{1}{N}\sum_{n=1}^{N} 1_{[t_n=k]}$$

$$p(t = \text{“5”}) = p(t = \text{“6”}) = 0.5$$

Testing with New Samples

$$\arg\max_t p(t|\mathbf{x}_{te}) = \arg\max_t p(\mathbf{x}_{te}|t)p(t)$$

$$p(\mathbf{x}_{te}|t = \text{“5”})p(t = \text{“5”}) \quad \textbf{vs.} \quad p(\mathbf{x}_{te}|t = \text{“6”})p(t = \text{“6”})$$



Summary

- Probabilistic generative models
  - Model likelihood (class-conditional density) p(x|t) and class prior p(t)
  - Maximum likelihood estimation on joint distribution p(x, t)
  - Prediction: argmax p(t|x) <=> argmax p(x|t) p(t)
- Gaussian Discriminant Analysis
  - Continuous random variables
  - p(x|t): Gaussian distributions with shared covariance matrix
  - p(t): Bernoulli (Multinomial) distribution for binary (multi) class
- Naïve Bayes Classifier
  - Discrete random variables
  - Naïve Bayes assumption: p(x_i, x_j | t) = p(x_i | t) p(x_j | t), for all i,j
  - p(x_i|t) : Bernoulli distribution for binary features
  - p(t): Bernoulli (Multinomial) distribution for binary (multi) class

Comparison

Generative models

Learn the model that generates the observed data; 'Everything' are about the distribution

Pros

- Need less data due to prior knowledge of dist.
- Can work with missing data
- Can capture uncertainty, priors, etc
- Low variance

Cons

- High bias (underfitting): due to the generalization by assumed distribution
- Less accurate than discriminative models (when the distribution assumption is violated)

Discriminative models

Directly discriminative the observed data regardless of the underlying generation process

Pros

- Easy to model
- Low bias
- More accurate than generative models

Cons

- Need more data to avoid underfitting
- Not easy to work with missing data
- Cannot capture priors, uncertainty, etc.
- High variance (easy to overfitting)