Engineering Applications of Machine Learning and Data Analytics

Gregory Ditzler

Dept. of Electrical & Computer Engineering
ditzler@email.arizona.edu

Lecture Set Overview

- Generating Data
- Decision Making with Bayes
- Assessing Risk in a Prediction
- Reading: Chapter 3

Generating Data

Generating Data

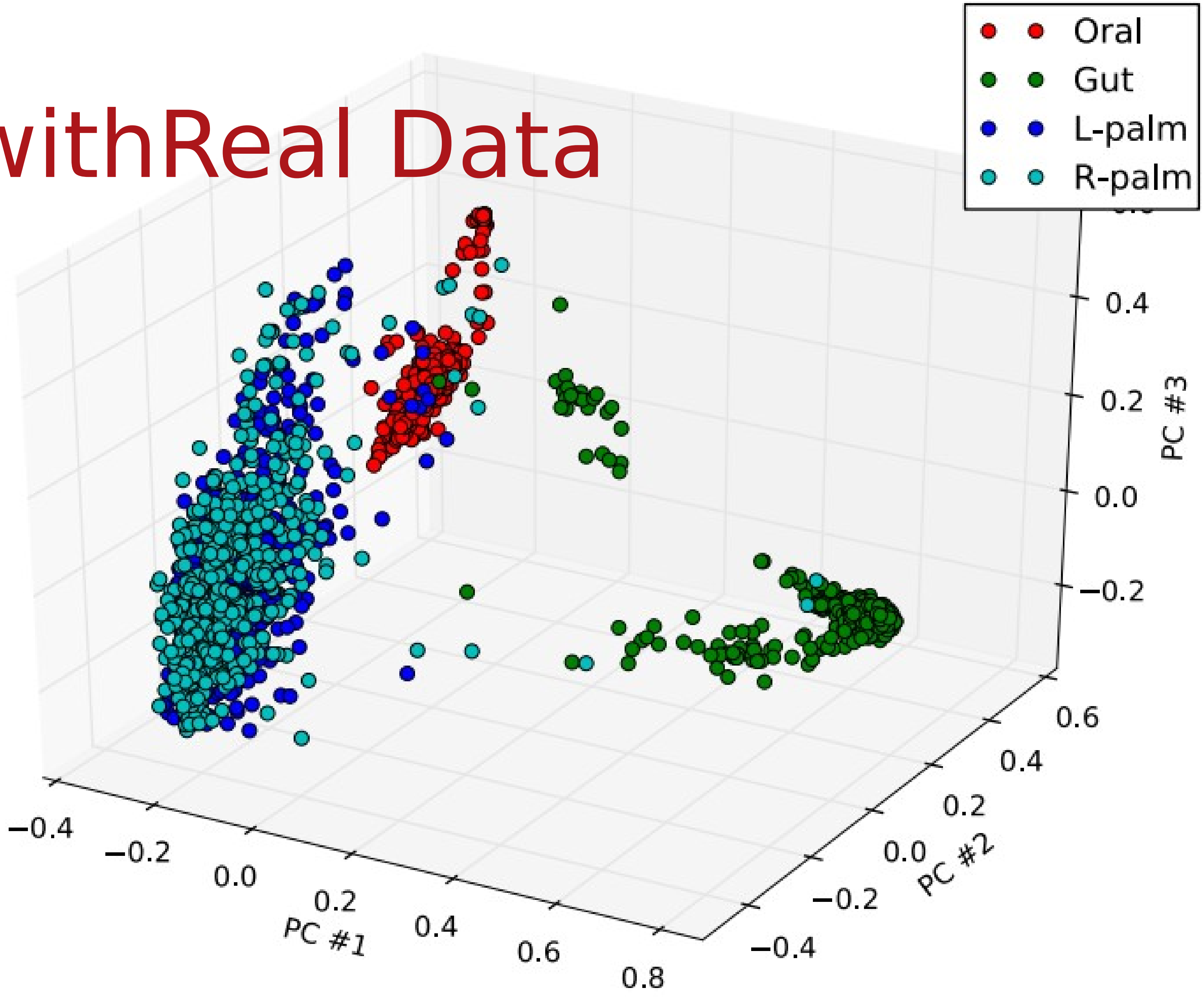
$X \leftarrow \text{measurements/variables}$ $\qquad Y \leftarrow \text{labels/classes}$

Generating Data

$X \leftarrow \text{measurements/variables} \qquad Y \leftarrow \text{labels/classes}$

- One aspect will encounter in machine learning is benchmarking our algorithms and the data we use in the experiment are very important
  - Real-world data are the true test to a model's performance but obtaining such data can be hard and expensive
  - Synthetic data can provide us a way to "*know*" the ground

Example withReal Data



Generating Data

$X \leftarrow \text{measurements/variables}$ $\qquad$ $Y \leftarrow \text{labels/classes}$

- One aspect will encounter in machine learning is benchmarking our algorithms and the data we use in the experiment are very important
  - Real-world data are the true test to a model's performance but obtaining such data can be hard and expensive
  - Synthetic data can provide us a way to "*know*" the ground truth for data
- How should we generate data if we want to do a simple classification problem, what distribution should we use?
  - What if I want to calculate $p(X)$, $p(Y)$ or $p(X|Y)$

Gaussian Distribution

- Generating multi-variate Gaussian data is one of the most popular ways to generate data.
- Let $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ then the probability distribution function is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)\right]$$

  - where $|\cdot|$ is the determinate of a matrix

Gaussian Distribution

- Generating multi-variate Gaussian data is one of the most popular ways to generate data.
- Let $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ then the probability distribution function is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)\right]$$

  - where $|\cdot|$ is the determinate of a matrix
- Gaussian data are common distribution to generate data

Gaussian Distribution

- Generating multi-variate Gaussian data is one of the most popular ways to generate data.
- Let $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ then the probability distribution function is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x} - \mu)\right]$$

  - where $|\cdot|$ is the determinate of a matrix
- Gaussian data are common distribution to generate data
- What happens if I want $p(X|Y)$?

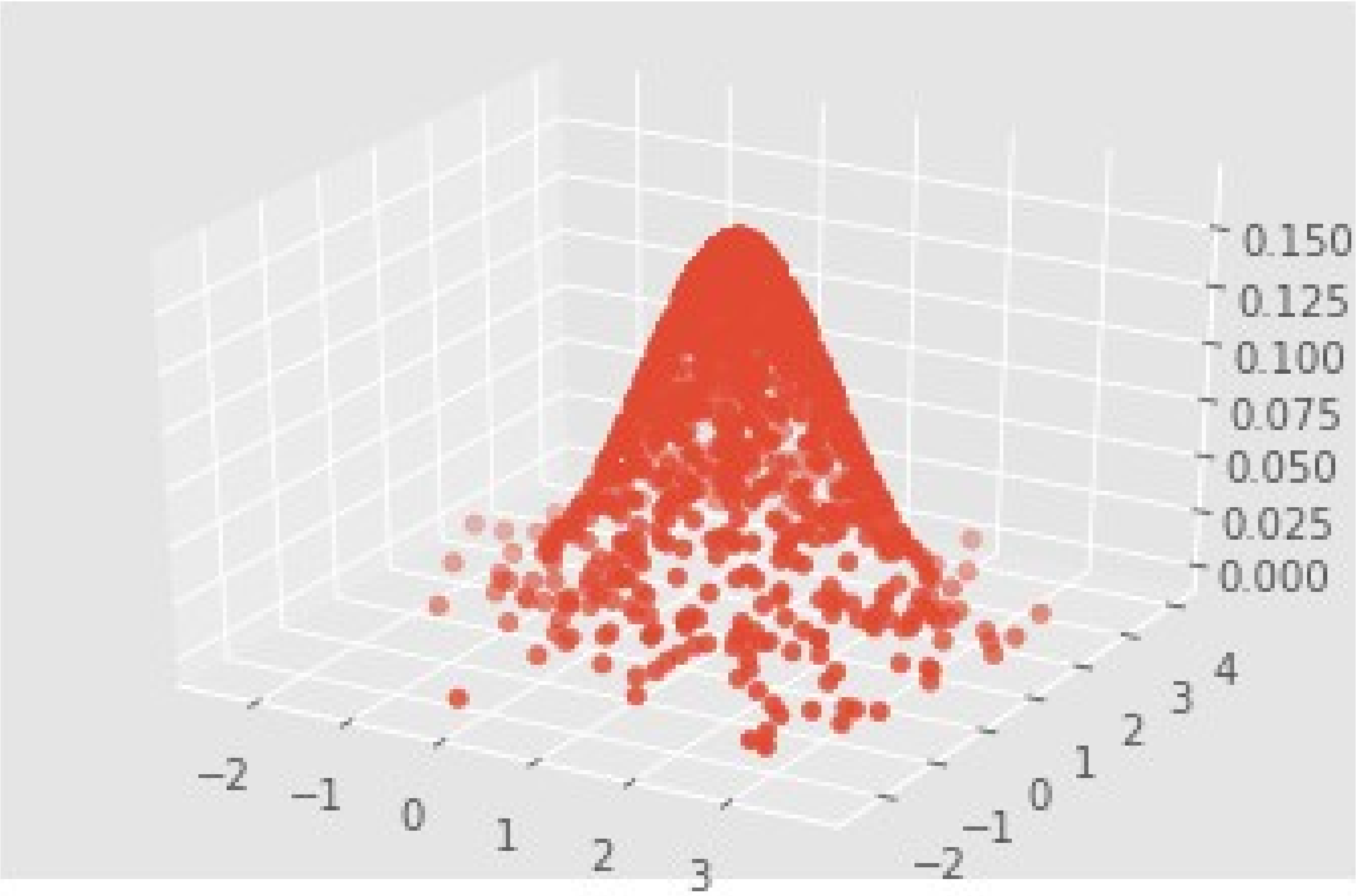
What if I wa$p(X|Y)$?

- For a two class problem where each class has a difference mean and variance then we have

$$p(\mathbf{x}|Y_1) = \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)^\mathsf{T}\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right]$$

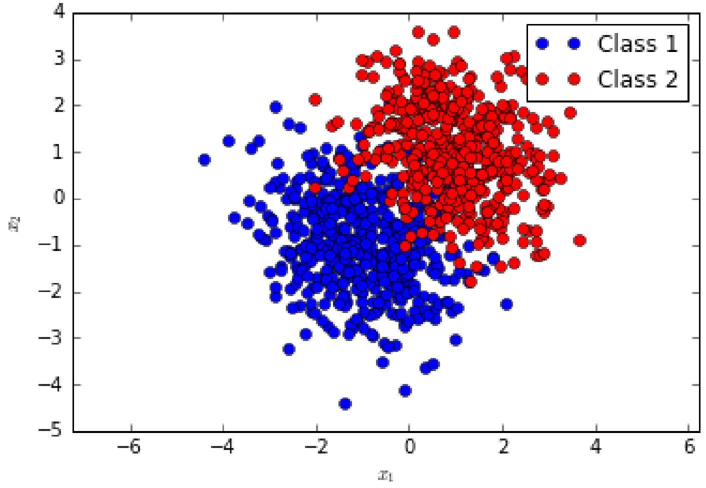
$$p(\mathbf{x}|Y_2) = \frac{1}{(2\pi)^{d/2}|\Sigma_2|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_2)^\mathsf{T}\Sigma_2^{-1}(\mathbf{x}-\mu_2)\right]$$

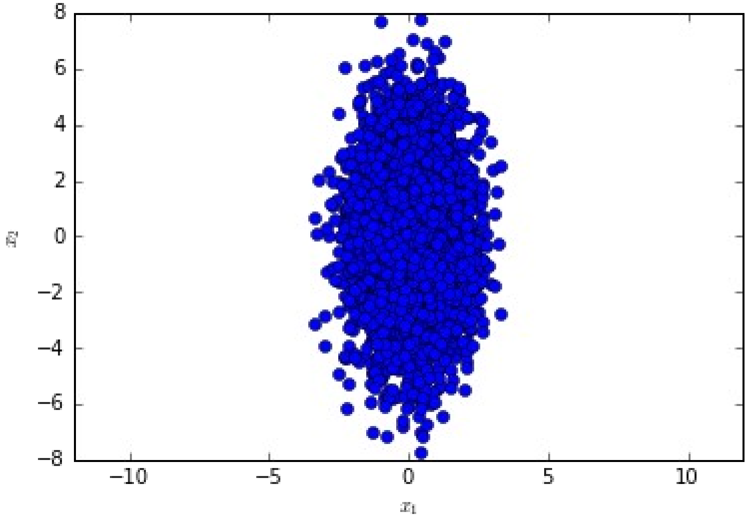
Visualizing the PDF of a 2D Gaussian

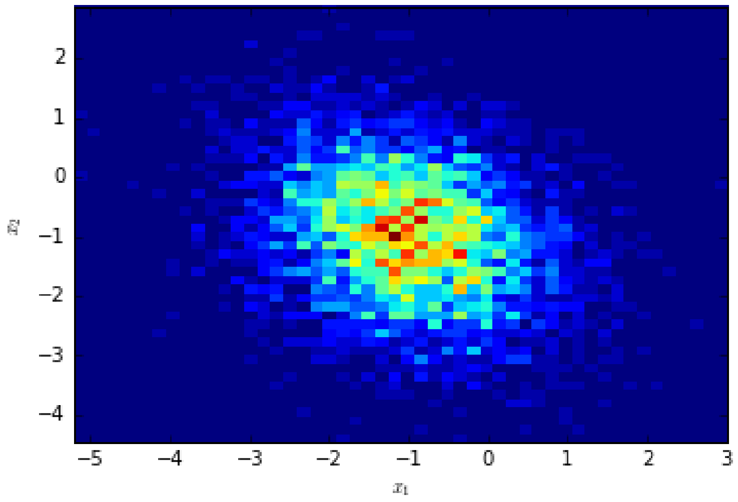


Generating Data from Probability Distributions

```
x = np.random.multivariate_normal([-1, -1], [[1, -.25], [-.25, 1]], 500).T
y = np.random.multivariate_normal([1, 1], [[1, -.25], [-.25, 1]], 500).T
plt.plot(x[0], x[1], 'o', c='b')
plt.plot(y[0], y[1], 'o', c='r')
plt.axis('equal')
plt.xlabel('$x_1$')  # use latex in the figure axis labels
plt.ylabel('$x_2$')
plt.legend(("Class 1", "Class 2"))
plt.show()
```



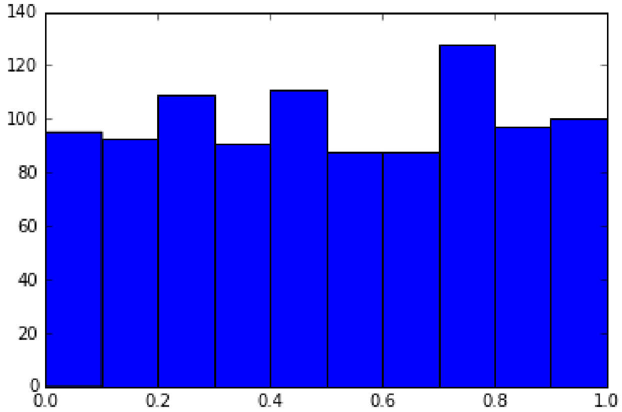




Generating Data from Probability Distributions

Uniform

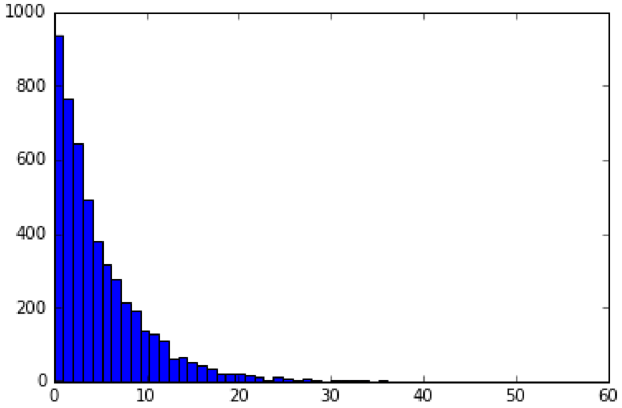
```
x = np.random.rand(1000)
d = plt.hist(x)
print "The mean of x is ", np.mean(x)
```



Exponential

$$f(x;\beta) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right)$$

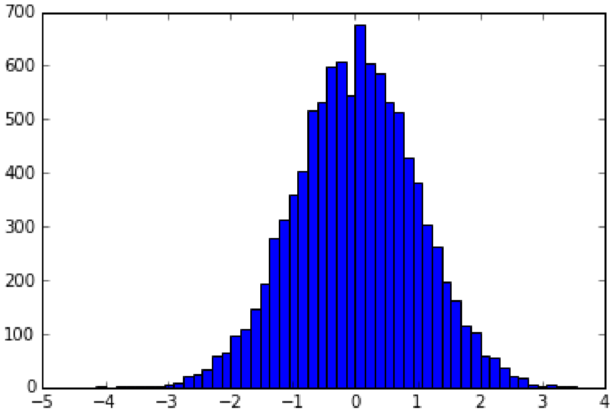
```
beta = .2
x = np.random.exponential(1/beta, 5000)
d = plt.hist(x, bins = 50)
print "The mean of x is ", np.mean(x)
```



Gaussian

$$f(x;\mu;\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)$$

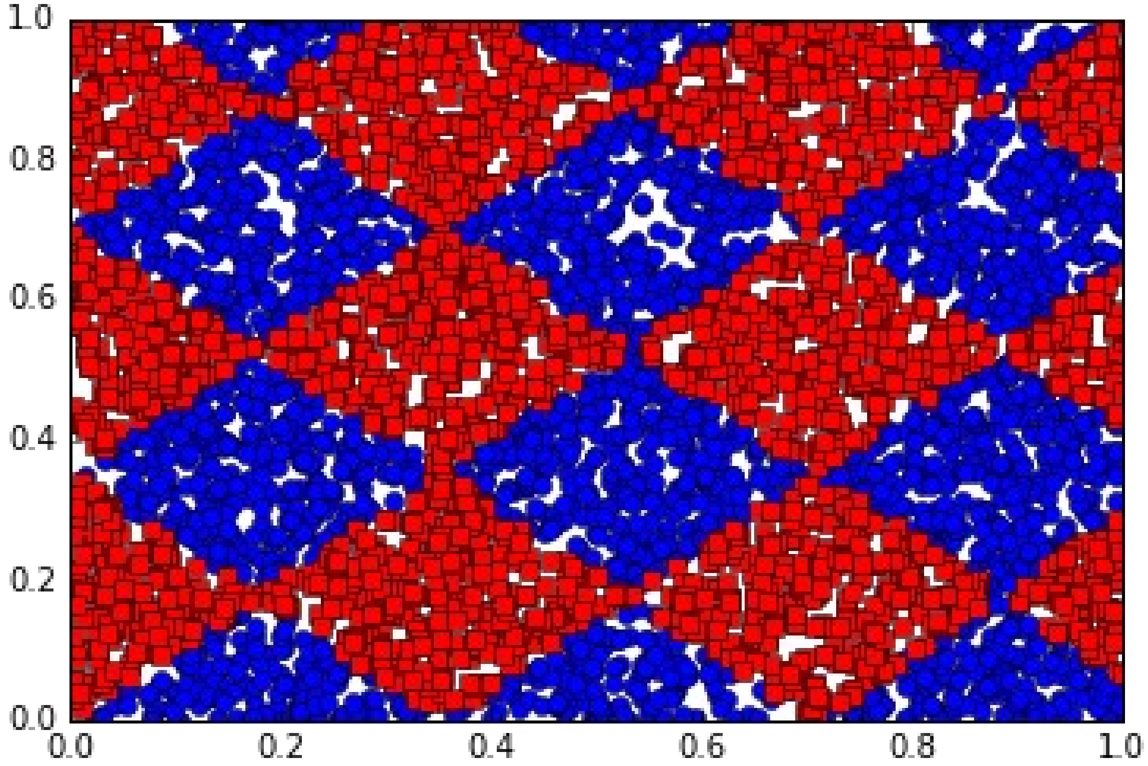
```
x = np.random.randn(10000)
d = plt.hist(x, bins=50)
print "The mean of x is ", np.mean(x)
```



Generating a Checkerboard

In Code

```
def gen_cb(N, a, alpha):
    """
    N: number of points on the checkerboard
    a: width of the checker board (0<a<1)
    alpha: rotation of the checkerboard in radians
    """
    d = np.random.rand(N, 2).T
    d_transformed = np.array([d[0]*np.cos(alpha)-d[1]*np.sin(alpha),
                              d[0]*np.sin(alpha)+d[1]*np.cos(alpha)]).T
    s = np.ceil(d_transformed[:,0]/a)+np.floor(d_transformed[:,1]/a)
    lab = 2 - (s%2)
    data = d.T
    return data, lab
```




```
X, y = gen_cb(5000, .25, 3.14159/4)
plt.figure()
plt.plot(X[np.where(y==1)[0], 0], X[np.where(y==1)[0], 1], 'o')
plt.plot(X[np.where(y==2)[0], 0], X[np.where(y==2)[0], 1], 's', c = 'r')
```

Distance

Distances: With and Without Distributions

- Many times throughout the semester we will need to either: (a) measure the size of a vector, or (b) measure the distance between two vectors

Distances: With and Without Distributions

- Many times throughout the semester we will need to either: (a) measure the size of a vector, or (b) measure the distance between two vectors
- Let us first consider a v $\mathbf{x} \in \mathbb{R}^p$ wit xi where i ∈ [p] := {1, . . . , p}. The r-norm given by

$$L_r(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i^r \right)^{\frac{1}{r}}$$

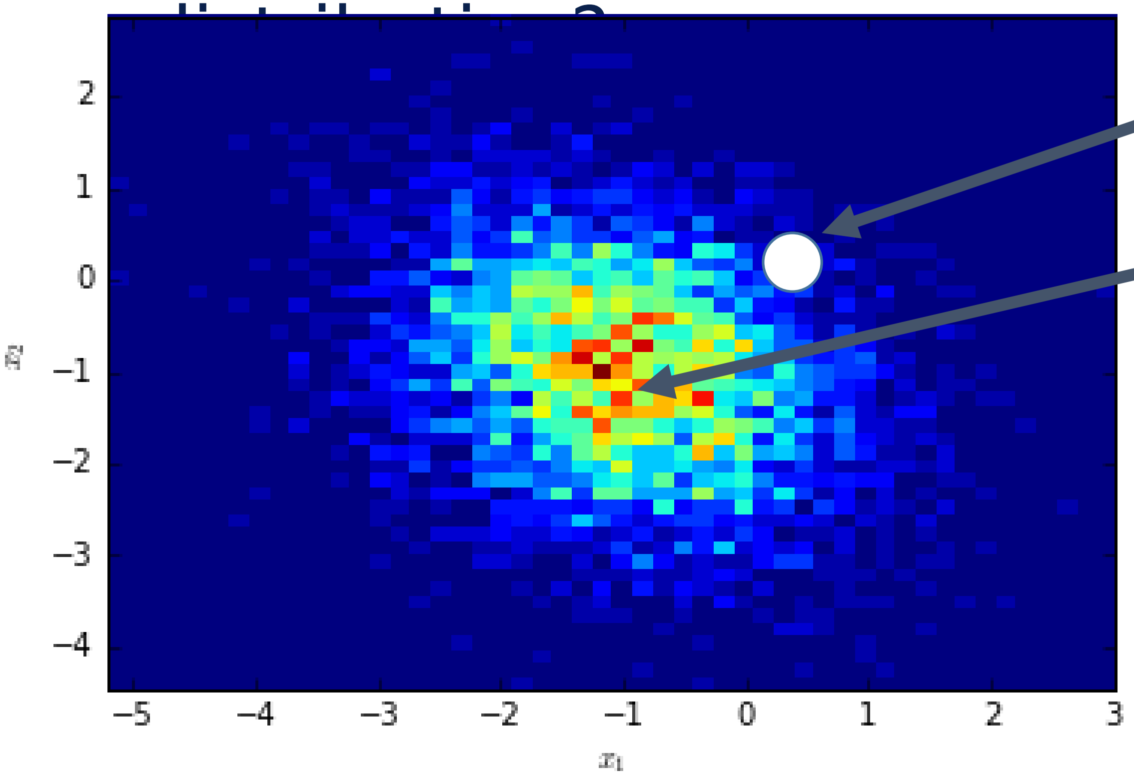
Distances: With and Without Distributions

- Many times throughout the semester we will need to either: (a) measure the size of a vector, or (b) measure the distance between two vectors
- Let us first consider a v$\mathbf{x} \in \mathbb{R}^p$ wit xi where i ∈ [p] := {1, . . . , p}. The r-norm given by $$L_r(\mathbf{x}) = \left(\sum_{i=1}^{p} |x_i|^r\right)^{\frac{1}{r}}$$
- This formulation can also be generalized to distances, which the most popular being the Euclidean Distance.

$$d_2(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{d} (x_i - z_i)^2\right)^{\frac{1}{2}}$$

The Distance Between a Point and Distribution

- Those definitions lend themselves well to measuring the distance between two points; however, what if we want to know the “distance” from a point **x** to a probability

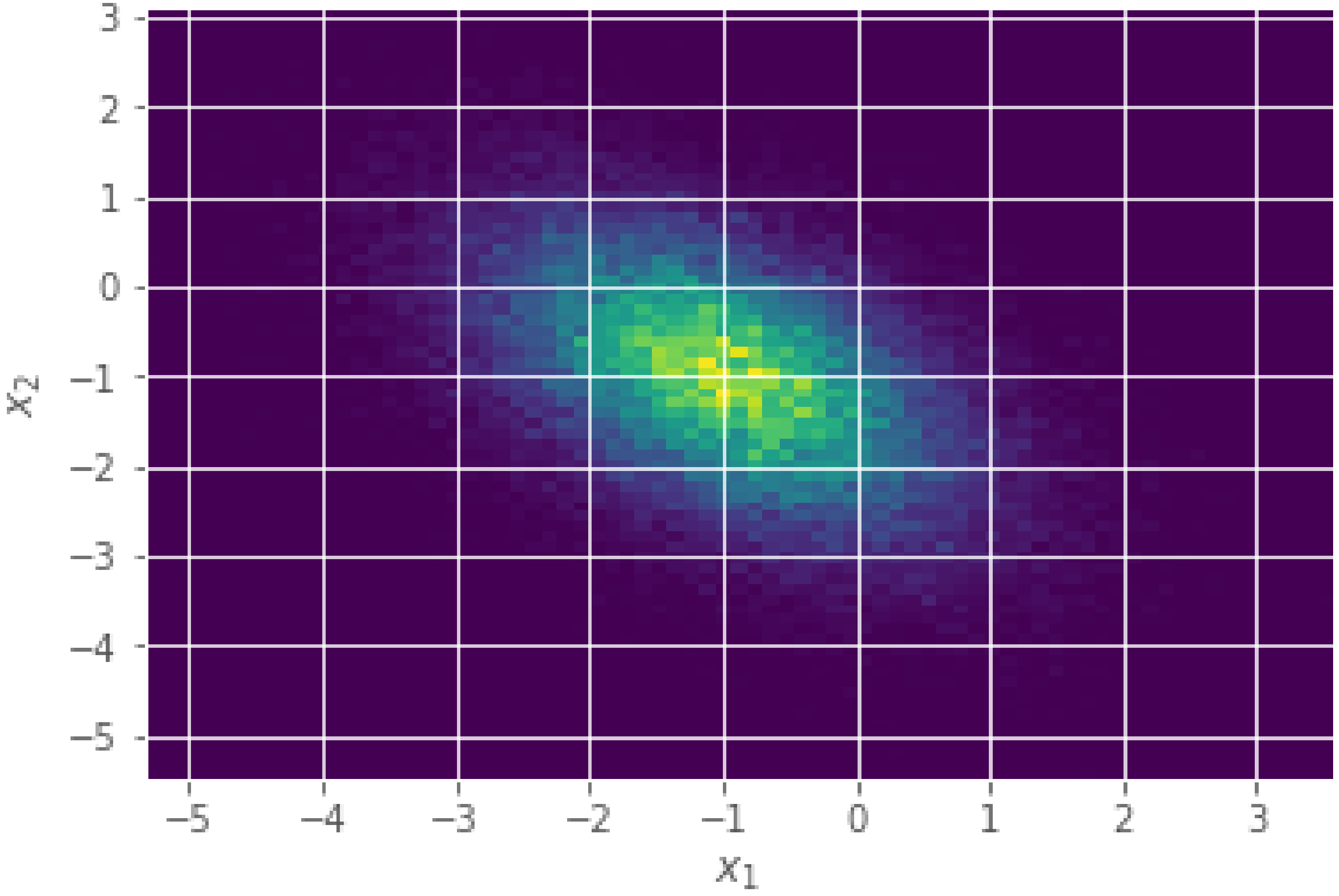


That point

That distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right]$$

Thinking about it



Mahalanobis Distance

- The Gaussian PDF has this distance hidden in the density function. The Mahalanobis Distance is the distance from a point to a distribution. The formal definition of this distance is given by

$$d_{\mathrm{Mahal}}(\mathbf{x}, \mu, \Sigma) = (\mathbf{x} - \mu)^{\mathsf{T}} \Sigma^{-1} (\mathbf{x} - \mu)$$

  - Note that a point x with "closer" Euclidean distance does not imply a closer Mahalanobis distance with a known μ and Σ

Probability & Bayes

Defining a Probability Distribution

- ***Sample space (Ω):*** the set of possible outcomes for the random experiment.
- ***Events (F):*** the set of possible events for the random experiment.
- ***Probability (P):*** a mass function $P : F \to [0,1]$ assigning a number $P(A) \in [0,1]$ to each $A \in F$, satisfying the axioms of probability given later in this section.
- ***Laws of Probability***
  - Non-negativity: $P(A) \geq 0$ for all $A \in F$.
  - Additivity: if A, B are disjoint events ($A \cap B = \emptyset$) then $P(A \cup B) = P(A) + P(B)$. More generally if $|\Omega| = \inf$ and $\{A_1, A_2, \ldots\}$ is an infinite sequence of disjoint events, then $P(A1 \cup A2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.
  - Normalization: $P(\Omega) = 1$ (something must happen)

Sum Rule

- The marginal probability of a single random variable can always be obtained by summing (integrating) the probability density function (pdf) over all values of all other variables. For example,

$$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$$

$$P(X) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P(X, Y = y, Z = z)$$

Product Rule

- The joint probability can always be obtained by multiplying the conditional probability (conditioned on one of the variables) with the marginal probability of the conditioned variable.

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

Bayes Rule

- The combination of these two rules gives us Bayes Rule. Let Y denote the outcome that we seek to predict (e.g., healthy or not) and X denote the variable(s) we are provided (e.g.,

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

Bayes Rule

- The combination of these two rules gives us Bayes Rule. Let Y denote the outcome that we seek to predict (e.g., healthy or not) and X denote the variable(s) we are provided (e.g., medical records).

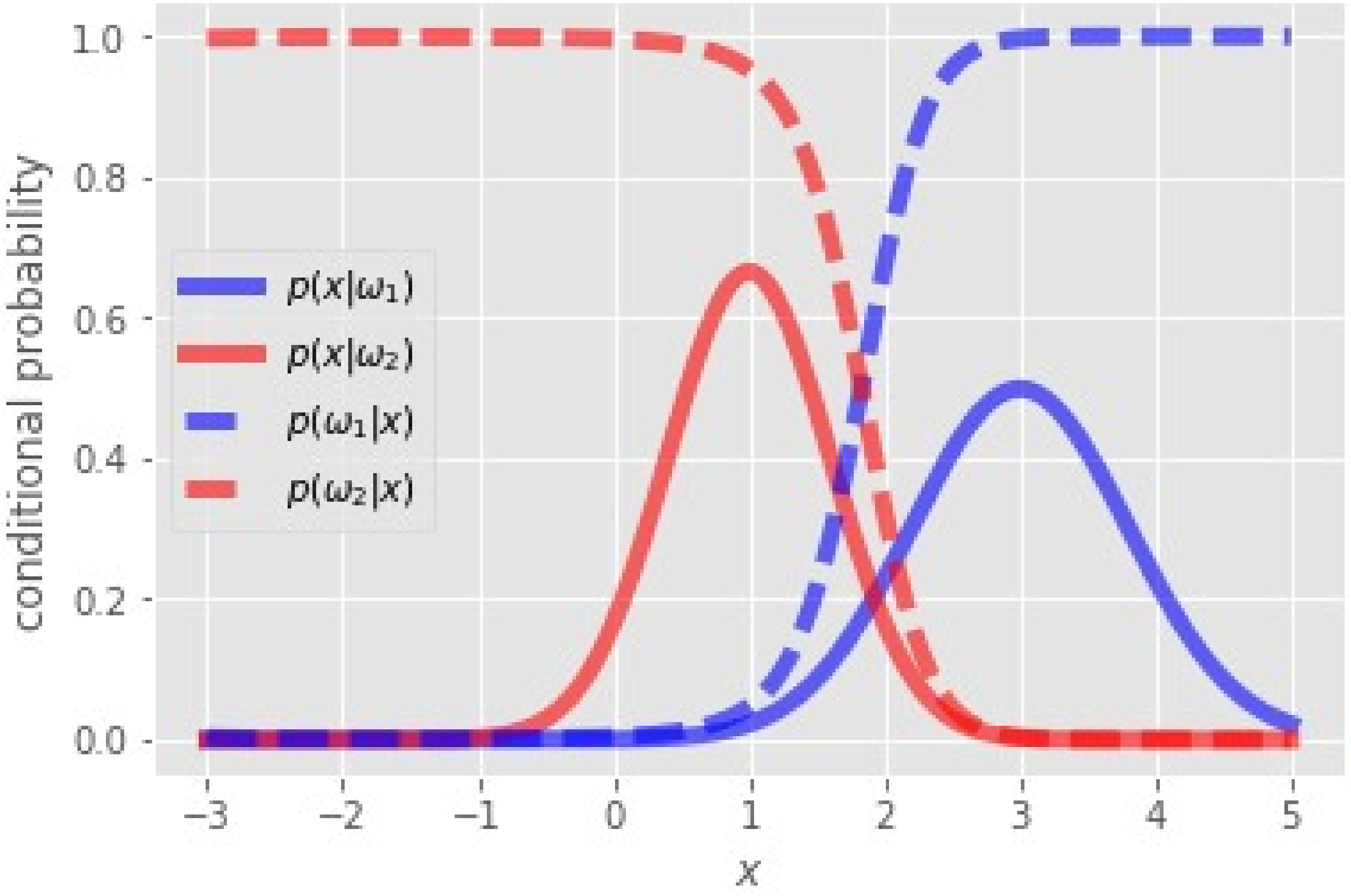
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_{y \in \mathcal{Y}} P(X, Y = y)}$$

Bayes Rule

- The combination of these two rules gives us Bayes Rule. Let Y denote the outcome that we seek to predict (e.g., healthy or not) and X denote the variable(s) we are provided (e.g., medical records).

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_{y \in \mathcal{Y}} P(X, Y = y)} = \frac{P(X|Y)P(Y)}{\sum_{y \in \mathcal{Y}} P(X|Y = y)P(Y = y)}$$

Visualizing the Posterior and Likelihood

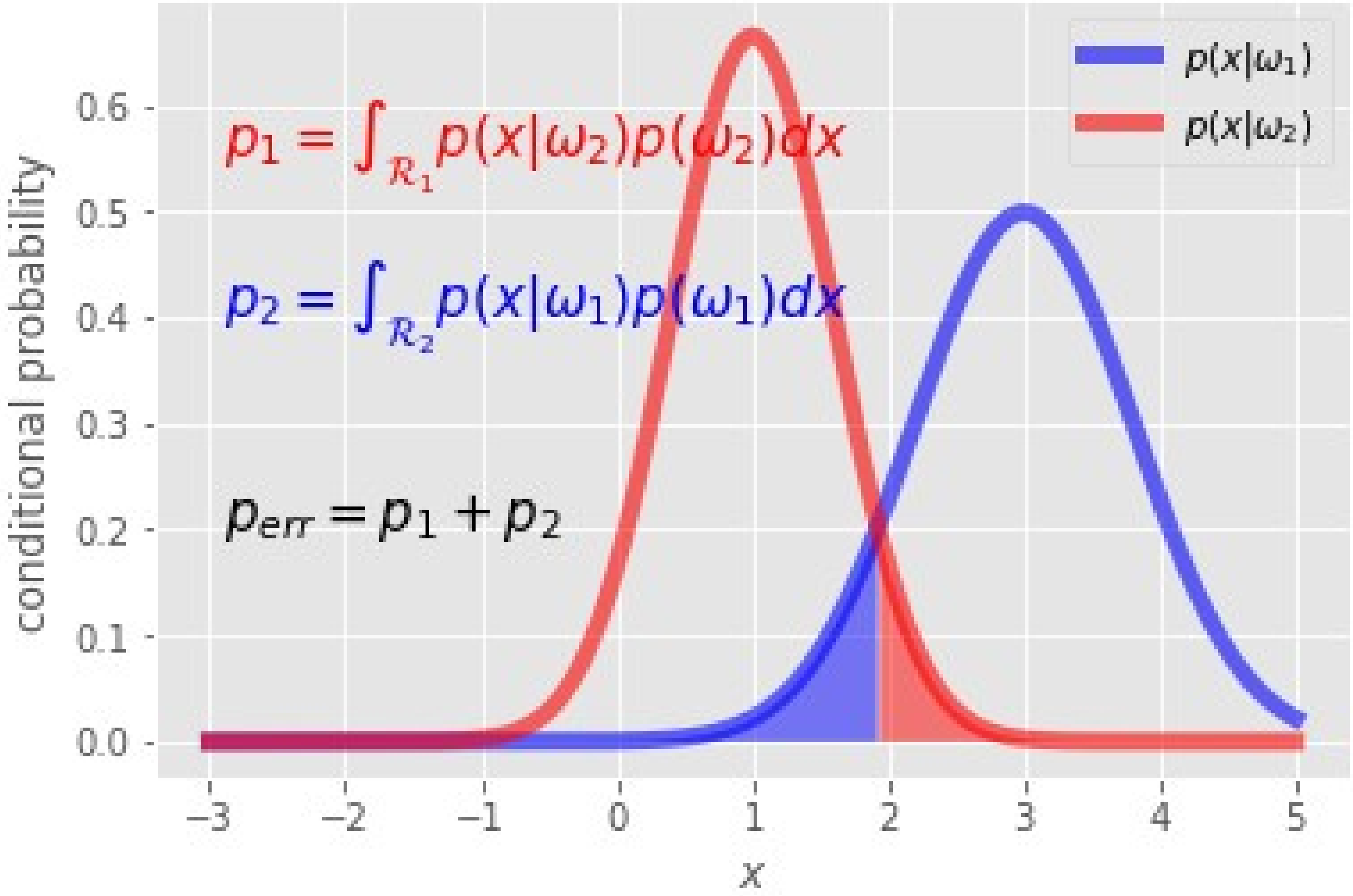


Decisions with Bayes Rule

- The Bayes decision rule is the one that *minimizes the probability of error* or the risk we take when a decision is made.

$$\omega^* = \arg\max_{\omega \in \Omega} P(\omega|X) = \arg\max_{\omega \in \Omega} \frac{P(X|\omega)P(\omega)}{P(X)} = \arg\max_{\omega \in \Omega} P(X|\omega)P(\omega)$$

Visualizing the Probability of Error



Example

My parents have two kids now grown into adults. Obviously there is me, Greg. I was born on a Wednesday. *What is the probability that I have a brother*? You can assume that P(boy) = P(girl) = 1/2.

Estimating the distributions

- What if we do not know the form of the distribution or we cannot determine a closed form?
- The simplest estimator of a probability mass function is a histogram.
  - That is to say if we have a two class problem the we can split the data into each class, we can attempt to model $p(x|y)$

Estimating the distributions

- What if we do not know the form of the distribution or we cannot determine a closed form?
- The simplest estimator of a probability mass function is a histogram.
  - That is to say if we have a two class problem the we can split the data into each class, we can attempt to model p(x|y) empirically using the sample that we have.

How much data is enough?

Example

Let us consider that our advisor told us that we need 30 samples in each bin of the histogram and there are 20 bins.

Example

Let us consider that our advisor told us that we need 30 samples in each bin of the histogram and there are 20 bins.

- That means for 1D data: $20 \times 30 = 600$ instances
- That means for 2D data: $20 \times 20 \times 30 = 12k$ instances
- That means for 3D data: $20 \times 20 \times 20 \times 30 = 240k$ instances

Meet the curse of dimensionality

Naïve Bayes Rule

- Estimating the likelihood terms can be extremely burdensome if we do not know the form of the distribution
- **Solution**: Naïve Bayes assumption. All the features are conditional

$$P(\omega)p(\mathbf{x}|\omega) = P(\omega)\prod_{i=1}^{p} p(x_i|\omega)$$

- The ma

$$P(\omega_j) = \frac{\#\text{of instances in } \omega_j}{\#\text{of instances in the data set}}$$

Naïve Bayes in Code

```
from sklearn import datasets
iris = datasets.load_iris()
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
y_pred = gnb.fit(iris.data, iris.target).predict(iris.data)
print("Number of mislabeled points out of a total %d points : %d"%
      (iris.data.shape[0],(iris.target != y_pred).sum()))
```

Risk in a Decision

Risk in a Decision

- Let us consider the possible class outcomes and the actions that we could consider to even make a decision
  - We have the class, or the state of nature, for an instance
  - We have the action that we (the classifier) took
  - We have a cost with making a decision
  - The number of actions we could take need not equal the number of classes
    - An action could be to not take an action

1/18/2017

Risk with Bayes

(HW #1 due 1/24)

classes: $\{\omega_1, \omega_2 \cdots \omega_c\}$

actions: $(\alpha_1, \alpha_2, \cdots, \alpha_\ell\}$

cost: $\{\lambda_1, \lambda_2 \cdots \lambda_\ell\}$

cost function: $\lambda(\alpha_i | \omega_j)$

"ℓ" need not equal "c"

Cost of taking action i given the state of nature is j

Risk in a Decision

$$\alpha = \arg\min_{\alpha_i} R(\alpha_i|\mathbf{x})$$

$$= \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

Conditional Risk with Bayes

$$\omega^* = \arg\max_{\omega \in \Omega} P(\omega|X)$$

$$R(\alpha_i|\vec{X}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|X)$$

$$\begin{array}{c|cc} & \multicolumn{2}{c}{\text{Predicted}} \\ & \omega_1 & \omega_2 \\ \hline \omega_1 & \lambda_{11} & \lambda_{12} \\ \omega_2 & \lambda_{21} & \lambda_{22} \end{array} \Bigg\} \text{True}$$

Bayes Rule for Last week

$$\lambda = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \{0\text{-}1 \text{ loss}\}$$

Example

$$\begin{array}{ccc} & \text{cancer} & \text{healthy} \\ \lambda = & \begin{bmatrix} 1/2 & 1000 \\ 10 & 0 \end{bmatrix} & \begin{array}{c} \text{cancer} \\ \text{healthy} \end{array} \end{array}$$

Risk in a Two Outcome Decision

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})$$

Lecture Set Overview

- Generating Data
- Decision Making with Bayes
- Assessing Risk in a Prediction
- Reading: Chapter 3