Machine Learning

4771

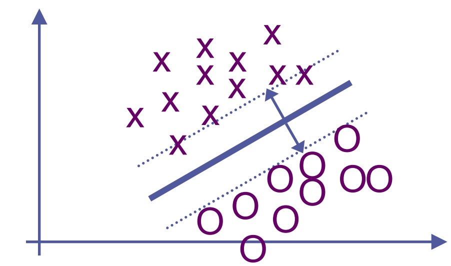
Instructor: Tony Jebara

Topic 7

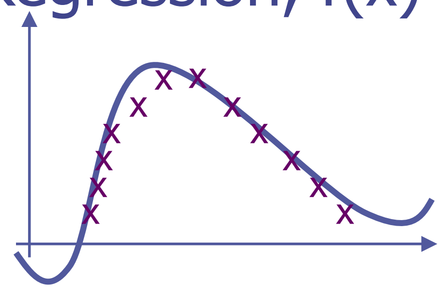
- Unsupervised Learning
- Statistical Perspective
- Probability Models
- Discrete & Continuous: Gaussian, Bernoulli, Multinomial
- Maximum Likelihood → Logistic Regression
- Conditioning, Marginalizing, Bayes Rule, Expectations
- Classification, Regression, Detection
- Dependence/Independence
- Maximum Likelihood → Naïve Bayes

Unsupervised Learning

Classification

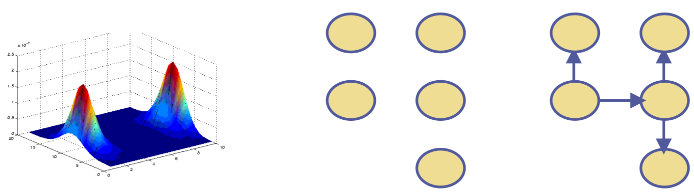


Regression, f(x)=y

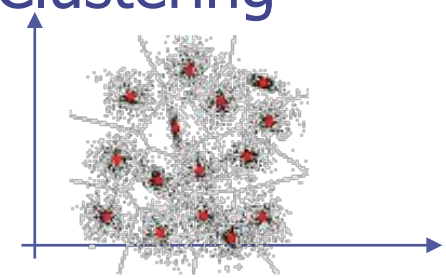


Supervised

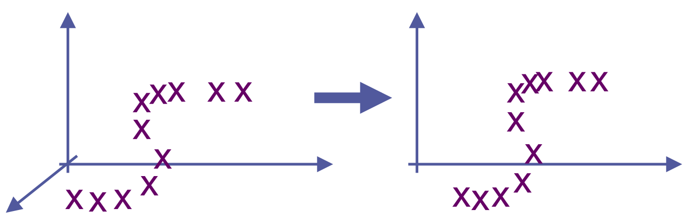
Density/Structure Estimation



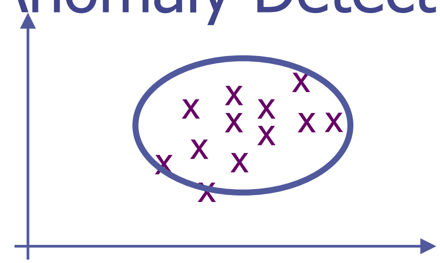
Clustering



Feature Selection



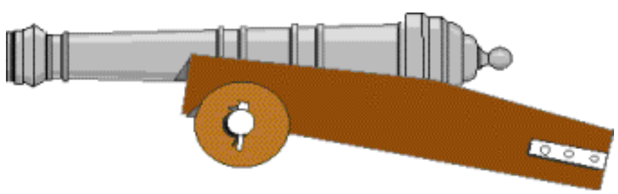
Anomaly Detection



Unsupervised
(can help
supervised)

Statistical Perspective

- Several problems with framework so far:
  - Only have input-output approaches (SVM, Neural Net)
  - Pulled non-linear squashing functions out of a hat
  - Pulled loss functions (squared error, etc.) out of a hat
- Better approach for classification?
- What if we have multi-class classification?
- What if other problems, i.e. unobserved values of x,y,etc…
- Also, what if we don't have a true function?
- Example of Projectile Cannon (c.f. Distal Learning)

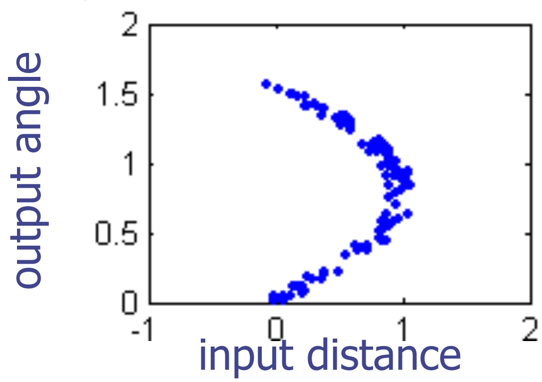


- Would like to train a regression function to control a cannon's angle of fire (y) given target distance (x)

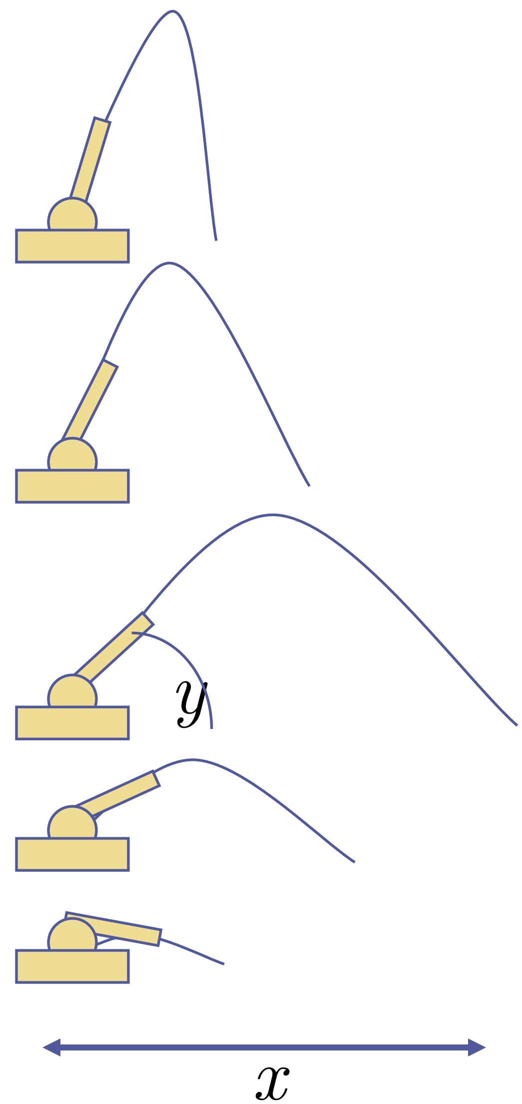
Statistical Perspective

- Example of Projectile Cannon
  (45 degree problem)
  x = input target distance
  y = output cannon angle

$$x = \frac{v(0)^2}{g}\sin(2y) + noise$$



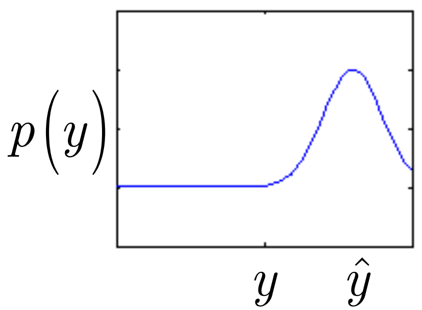
- What does least squares do?
- Conditional statistical models address this problem…



Probability Models

- Instead of deterministic functions, output is a probability
- Previously: our output was a scalar $\hat{y} = f\left(x\right) = \theta^T x + b$
- Now: our output is a probability $p\left(y\right)$

  e.g. a probability bump:

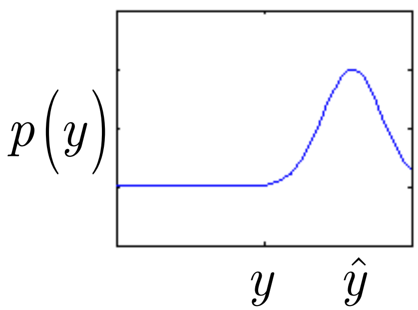


- $p\left(y\right)$ subsumes or is a superset of $\hat{y}$
- Why is this representation for our answer more general?

Probability Models

- Instead of deterministic functions, output is a probability
- Previously: our output was a scalar $\hat{y} = f(x) = \theta^T x + b$
- Now: our output is a probability $p(y)$

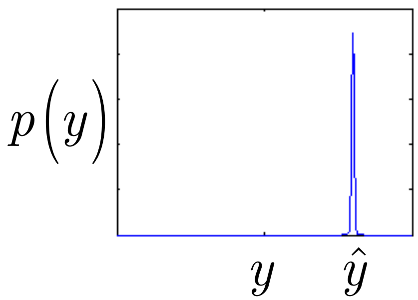
  e.g. a probability bump:



- $p(y)$ subsumes or is a superset of $\hat{y}$
- Why is this representation for our answer more general?

  → A deterministic answer $\hat{y}$ with complete confidence is like putting a probability $p(y)$ where all the mass is at $\hat{y}$!

$$\hat{y} \Leftrightarrow p(y) = \delta(y - \hat{y})$$



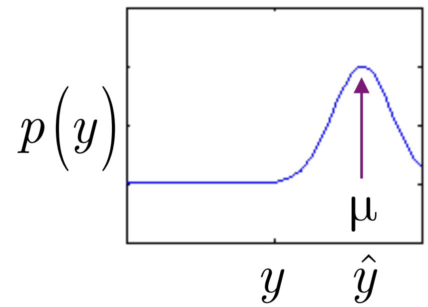
Probability Models

- Now: our output is a probability density function (pdf) $p(y)$
- Probability Model: a family of pdf's with adjustable parameters which lets us select one of many

$$p(y) \rightarrow p(y \mid \Theta)$$

- E.g.: 1-dim Gaussian distribution 'given' 'mean' parameter μ:

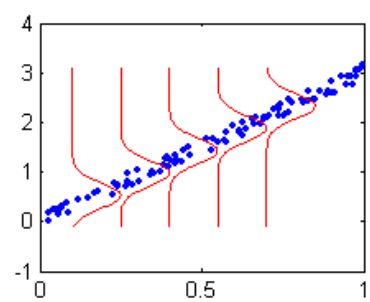
$$p(y \mid \mu) = N(y \mid \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$

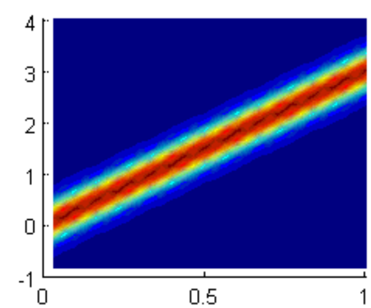


- Want mean centered on f(x)'s value $p(y) = N(y \mid f(x))$
- Now, linear regression is:

$$N(y \mid f(x)) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-f(x))^2}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta^T x-b)^2}$$





Probability Models

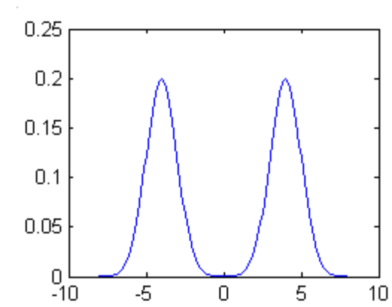
- To fit to data, we typically "maximize likelihood" of the probability model
- Log-likelihood = objective function (i.e. negative of cost) for probability models which we want to maximize
- Define (conditional) likelihood as $L(\Theta) = \prod_{i=1}^{N} p(y_i \mid x_i)$

  or log-Likelihood as $l(\Theta) = \log(L(\Theta)) = \sum_{i=1}^{N} \log p(y_i \mid x_i)$
- For Gaussian p(y|x), maximum likelihood is least squares!

$$\sum_{i=1}^{N} \log p(y_i \mid x_i) = \sum_{i=1}^{N} \log N(y_i \mid f(x_i)) = \sum_{i=1}^{N} \log \tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y_i - f(x_i)\right)^2}$$

$$= -N \log\left(\sqrt{2\pi}\right) - \sum_{i=1}^{N} \tfrac{1}{2}\left(y_i - f(x_i)\right)^2$$

Probability Models



- Can extend probability model to 2 bumps:

$$p\left(y \mid \Theta\right) = \tfrac{1}{2} N\left(y \mid \mu_1\right) + \tfrac{1}{2} N\left(y \mid \mu_2\right)$$

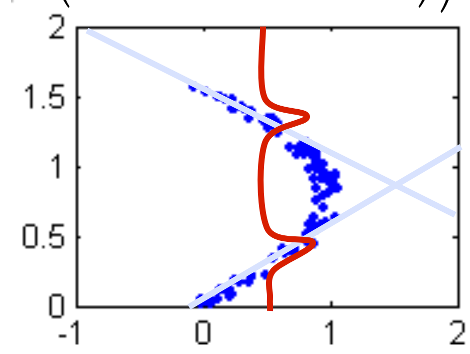
- Each mean can be a linear regression fn.

$$p\left(y \mid x, \Theta\right) = \tfrac{1}{2} N\left(y \mid f_1\left(x\right)\right) + \tfrac{1}{2} N\left(y \mid f_2\left(x\right)\right)$$

$$= \tfrac{1}{2} N\left(y \mid \theta_1^T x + b_1\right) + \tfrac{1}{2} N\left(y \mid \theta_2^T x + b_2\right)$$

- Therefore the (conditional) log-likelihood to maximize is:

$$l\left(\Theta\right) = \sum_{i=1}^{N} \log\left(\tfrac{1}{2} N\left(y_i \mid \theta_1^T x_i + b_1\right) + \tfrac{1}{2} N\left(y_i \mid \theta_2^T x_i + b_2\right)\right)$$



- Maximize $l(\theta)$ using gradient ascent
- Nicely handles the “cannon firing” data

Probability Models

- Now classification: can also go beyond deterministic!
- Previously: wanted output to be binary $\hat{y} = \{0,1\}$
- Now: our output is a probability $p(y)$

  e.g. a probability table:

y=0	y=1
0.73	0.27

α

- This subsumes or is a superset again…
- Consider probability over binary events (coin flips!):

  e.g. Bernoulli distribution (i.e 1x2 probability table)
  with parameter α

$$p(y \mid \alpha) = \alpha^{y} (1 - \alpha)^{1-y} \qquad \alpha \in [0,1]$$

- Linear classification can be done by setting α equal to f(x):

$$p(y \mid x) = f(x)^{y} \left(1 - f(x)\right)^{1-y} \qquad f(x) \in [0,1]$$

Probability Models

- Now linear classification is:

$$p\left(y \mid x\right) = f\left(x\right)^{y}\left(1 - f\left(x\right)\right)^{1-y} \qquad f\left(x\right) \equiv \alpha \in \left[0,1\right]$$

- Log-likelihood is (negative of cost function):

$$\begin{aligned} \sum_{i=1}^{N} \log p\left(y_i \mid x_i\right) &= \sum_{i=1}^{N} \log f\left(x_i\right)^{y_i}\left(1 - f\left(x_i\right)\right)^{1-y_i} \\ &= \sum_{i=1}^{N} y_i \log f\left(x_i\right) + \left(1 - y_i\right)\log\left(1 - f\left(x_i\right)\right) \\ &= \sum_{i \in class1} \log f\left(x_i\right) + \sum_{i \in class0} \log\left(1 - f\left(x_i\right)\right) \end{aligned}$$

- But, need a squashing function since f(x) in [0,1]
- Use sigmoid or logistic again…

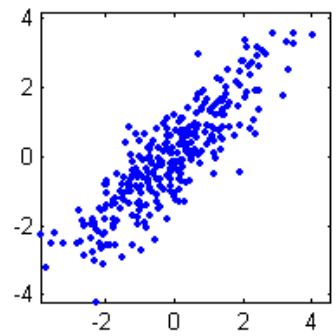
$$f\left(x\right) = sigmoid\left(\theta^T x + b\right) \in [0,1]$$

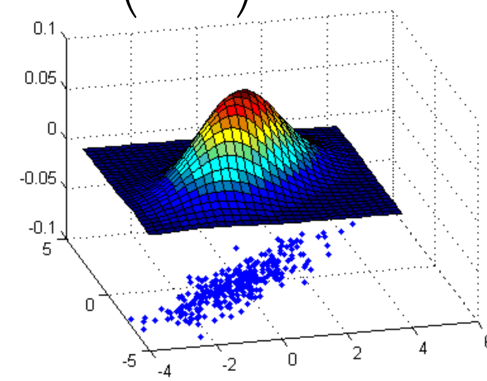
- Called logistic regression → *new loss function*
- Do gradient descent, similar to logistic output neural net!
- Can also handle multi-layer f(x) and do backprop again!

Generative Probability Models

- Idea: Extend probability to describe *both* X and Y
- Find probability density function over both: $p\left(x, y\right)$

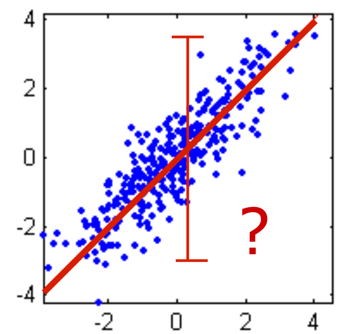
E.g. *describe* data with Multi-Dim. Gaussian (later…)





- Called a 'Generative Model' because we can use it to synthesize or re-generate data similar to the training data we learned from

- Regression models & classification boundaries are not as flexible don't keep info about X don't model noise/uncertainty



Properties of PDFs

- Let's review some basics of probability theory
- First, pdf is a function, multiple inputs, one output:

$$p\left(x_1,\ldots,x_n\right) \qquad\qquad p\left(x_1 = 0.3,\ldots,x_n = 1\right) = 0.2$$

- Function's output is always non-negative:

$$p\left(x_1,\ldots,x_n\right) \geq 0$$

- Can have discrete or continuous or both inputs:

$$p\left(x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 3.1415\right)$$

- Summing over the domain of all inputs gives unity:

$$\int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} p\left(x,y\right) dx dy = 1 \qquad \sum_y \sum_x p\left(x,y\right) = 1$$

0.4	0.1
0.3	0.2

Continuous→integral, Discrete→sum

Properties of PDFs

- Marginalizing: integrate/sum out a variable leaves a marginal distribution over the remaining ones…

$$\sum_y p(x,y) = p(x)$$

- Conditioning: if a variable 'y' is 'given' we get a conditional distribution over the remaining ones…

$$p(x \mid y) = \frac{p(x,y)}{p(y)}$$

- Bayes Rule: mathematically just redo conditioning but has a deeper meaning (1764)… if we have X being data and θ being a model

$$\underbrace{p(\theta \mid \mathcal{X})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{X} \mid \theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{X})}_{\text{evidence}}}$$

Properties of PDFs

- Expectation: can use pdf p(x) to compute averages and expected values for quantities, denoted by:

$$E_{p(x)}\left\{f\left(x\right)\right\} = \int_x p\left(x\right) f\left(x\right) dx \quad or \quad = \sum_x p\left(x\right) f\left(x\right)$$

- Properties:

$$E\left\{cf\left(x\right)\right\} = cE\left\{f\left(x\right)\right\}$$

$$E\left\{f\left(x\right) + c\right\} = E\left\{f\left(x\right)\right\} + c$$

$$E\left\{E\left\{f\left(x\right)\right\}\right\} = E\left\{f\left(x\right)\right\}$$

example: speeding ticket

Fine=0$	Fine=20$
0.8	0.2

expected cost of speeding?
f(x=0)=0, f(x=1)=20
p(x=0)=0.8, p(x=1)=0.2

- Mean: expected value for x

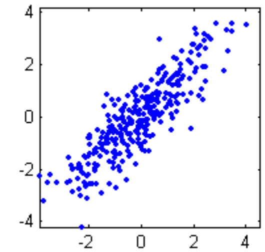
$$E_{p(x)}\left\{x\right\} = \int_{-\infty}^{\infty} p\left(x\right) x \, dx$$

- Variance: expected value of (x-mean)^2, how much x varies

$$Var\left\{x\right\} = E\left\{\left(x - E\left\{x\right\}\right)^2\right\} = E\left\{x^2 - 2xE\left\{x\right\} + E\left\{x\right\}^2\right\}$$

$$= E\left\{x^2\right\} - 2E\left\{x\right\}E\left\{x\right\} + E\left\{x\right\}^2 = E\left\{x^2\right\} - E\left\{x\right\}^2$$

Properties of PDFs



- Covariance: how strongly x and y vary together

$$Cov\left\{x,y\right\} = E\left\{\left(x - E\left\{x\right\}\right)\left(y - E\left\{y\right\}\right)\right\} = E\left\{xy\right\} - E\left\{x\right\}E\left\{y\right\}$$

- Conditional Expectation: $E\left\{y \mid x\right\} = \int_y p\left(y \mid x\right) y\, dy$

$$E\left\{E\left\{y \mid x\right\}\right\} = \int_x p\left(x\right)\int_y p\left(y \mid x\right) y\, dy\, dx = E\left\{y\right\}$$

- Sample Expectation: If we don't have pdf p(x,y) can approximate expectations using samples of data

$$E_{p(x)}\left\{f\left(x\right)\right\} \simeq \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right)$$

- Sample Mean: $E\left\{x\right\} \simeq \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

- Sample Var: $E\left\{\left(x - E\left(x\right)\right)^2\right\} \simeq \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$

- Sample Cov: $E\left\{\left(x - E\left(x\right)\right)\left(y - E\left(y\right)\right)\right\} \simeq \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$

More Properties of PDFs

- Independence: probabilities of independent variables multiply. Denote with the following notation:

$$x \perp\!\!\!\perp y \quad \rightarrow \quad p\left(x,y\right) = p\left(x\right)p\left(y\right)$$

$$x \perp\!\!\!\perp y \quad \rightarrow \quad p\left(x \mid y\right) = p\left(x\right)$$

also note in this case:

$$E_{p\left(x,y\right)}\left\{xy\right\} = \int_x \int_y p\left(x\right)p\left(y\right)xy\,dx\,dy$$

$$= \int_x p\left(x\right)x\,dx \int_y p\left(y\right)y\,dy = E_{p\left(x\right)}\left\{x\right\}E_{p\left(y\right)}\left\{y\right\}$$

- Conditional independence: when two variables become independent only if another is observed

$$x \perp\!\!\!\perp y \mid z \quad \rightarrow \quad p\left(x \mid y,z\right) = p\left(x \mid z\right)$$

$$x \perp\!\!\!\perp y \mid z \quad \rightarrow \quad p\left(x \mid y\right) \neq p\left(x\right)$$

The IID Assumption

- Most of the time, we will assume that a dataset independent and identically distributed (IID)

- In many real situations, data is generated by some black box phenomenon in an arbitrary order.
- Assume we are given a dataset:

$$\mathcal{X} = \left\{ x_1, \ldots, x_N \right\}$$

"Independent" means that (given the model θ) the probability of our data multiplies:

$$p\left(x_1, \ldots, x_N \mid \Theta\right) = \prod_{i=1}^{N} p_i\left(x_i \mid \Theta\right)$$

"Identically distributed" means that each marginal probability is the same for each data point

$$p\left(x_1, \ldots, x_N \mid \Theta\right) = \prod_{i=1}^{N} p_i\left(x_i \mid \Theta\right) = \prod_{i=1}^{N} p\left(x_i \mid \Theta\right)$$

The IID Assumption

- Bayes rule says likelihood is probability of data given model

$$\underbrace{p(\theta \mid \mathcal{X})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{X} \mid \theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{X})}_{\text{evidence}}}$$

- The likelihood of $\mathcal{X} = \{x_1, \ldots, x_N\}$ under IID assumptions is:

$$p(\mathcal{X} \mid \Theta) = p(x_1, \ldots, x_N \mid \Theta) = \prod_{i=1}^{N} p_i(x_i \mid \Theta) = \prod_{i=1}^{N} p(x_i \mid \Theta)$$

- Learn joint distribution $p(x \mid \Theta)$ by maximum likelihood:

$$\Theta^* = \arg\max_\Theta \prod_{i=1}^{N} p(x_i \mid \Theta) = \arg\max_\Theta \sum_{i=1}^{N} \log p(x_i \mid \Theta)$$

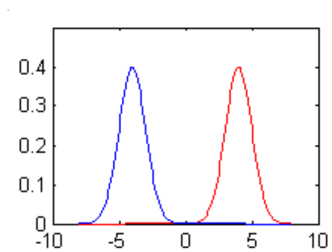
- Learn conditional $p(y \mid x, \Theta)$ by max conditional likelihood:

$$\Theta^* = \arg\max_\Theta \prod_{i=1}^{N} p(y_i \mid x_i, \Theta) = \arg\max_\Theta \sum_{i=1}^{N} \log p(y_i \mid x_i, \Theta)$$

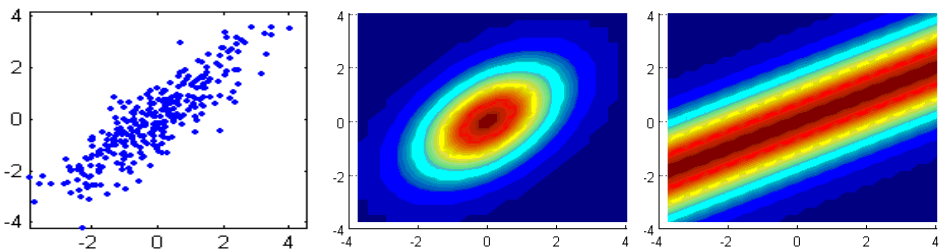
Uses of PDFs

- **Classification:** have p(x,y) and given x. Asked for discrete y output, give most probable one

  $$p\left(x,y\right) \rightarrow p\left(y \mid x\right) \rightarrow \hat{y} = \arg\max_{m} p\left(y = m \mid x\right)$$



- **Regression:** have p(x,y) and given x. Asked for a scalar y output, give most probable or expected one



$$\hat{y} = \begin{cases} \arg\max_{y} p\left(y \mid x\right) \\ E_{p\left(y \mid x\right)}\left\{y\right\} \end{cases}$$

$$\left\{\left(x_i, y_i\right)\right\} \rightarrow p\left(x,y\right) \rightarrow p\left(y \mid x\right)$$

- **Anomaly Detection:** if have p(x,y) and given both x,y. Asked if it is similar → threshold

  $$p\left(x,y\right) \geq threshold \rightarrow \left\{normal, anomaly\right\}$$