Machine Learning

4771

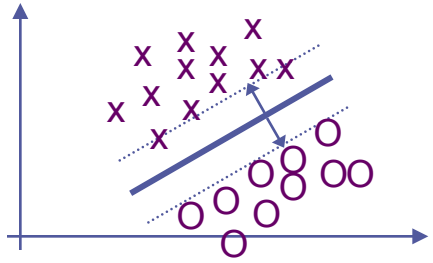
Instructor: Tony Jebara

Topic 7

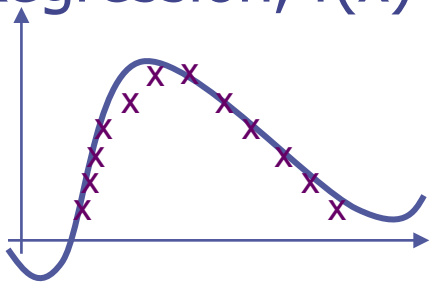
- Unsupervised Learning
- Statistical Perspective
- Probability Models
- Discrete & Continuous: Gaussian, Bernoulli, Multinomial
- Maximum Likelihood → Logistic Regression
- Conditioning, Marginalizing, Bayes Rule, Expectations
- Classification, Regression, Detection
- Dependence/Independence
- Maximum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

Unsupervised Learning

Classification

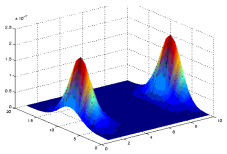


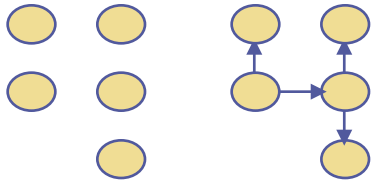
Regression, f(x)=y



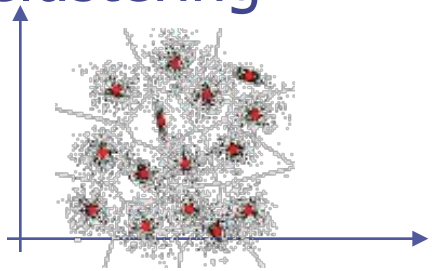
Supervised

Density/Structure Estimation



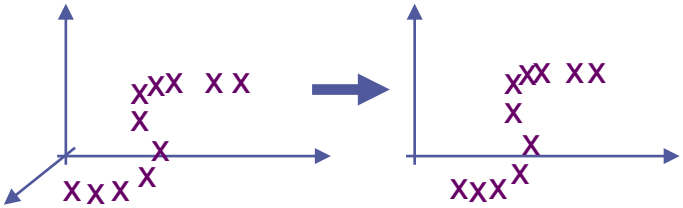


Clustering

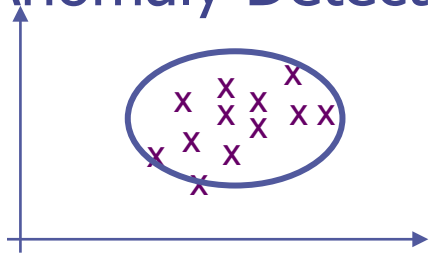


Unsupervised (can help supervised)

Feature Selection

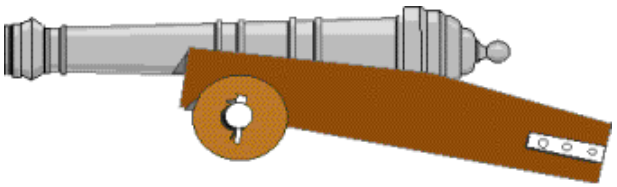


Anomaly Detection



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)

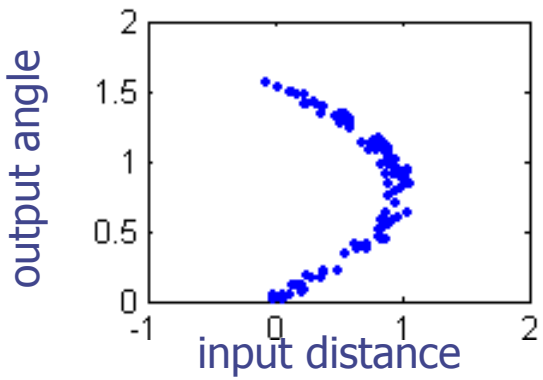


- Would like to train a regression function to control a cannon's angle of fire (y) given target distance (x)

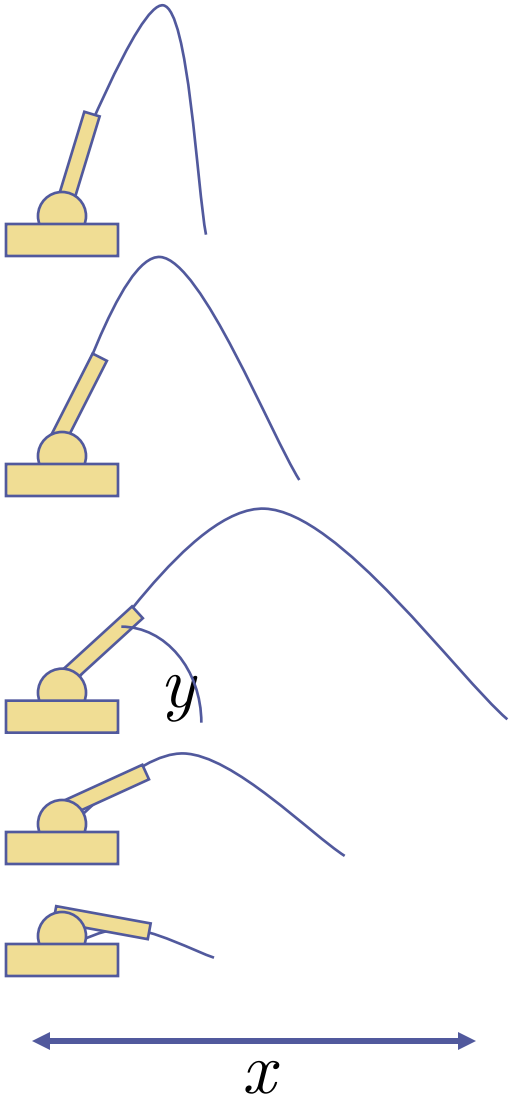
Statistical Perspective

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

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



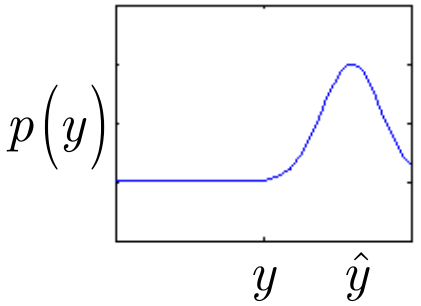
- What does least squares do?
- Conditional statistical models address this problem…



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:

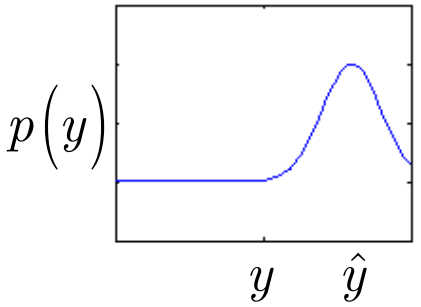


- $p\left(y\right)$ subsumes or is a superset of $\hat{y}$
- Why is this representation for our answer more general?

Probability Models

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

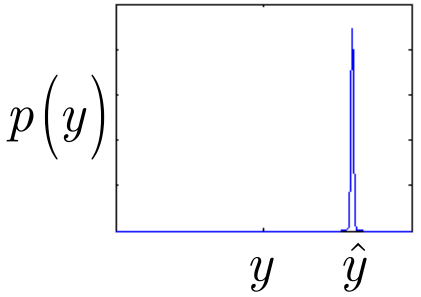
  e.g. a probability bump:



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

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

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



pdf: probability density function.

Probability Models

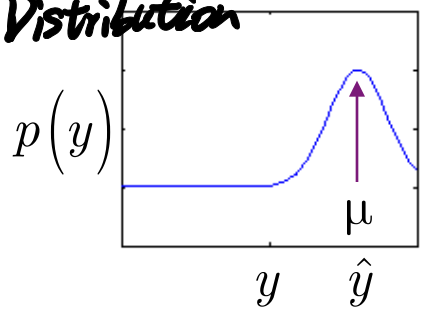
- Now: our output is a probability density function (pdf) $p(y)$
- Probability Model: a family of pdf's with adjustable parameters which lets us select one of many

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

Normal Distribution = Gaussian Distribution

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

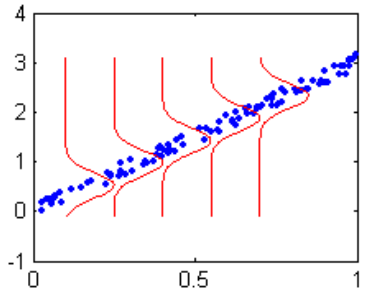
$$p(y \mid \mu) = N(y \mid \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$

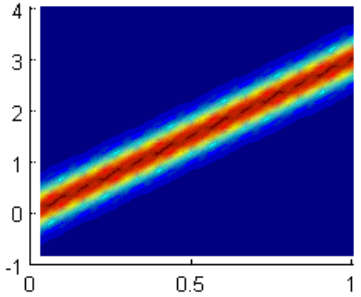


- Want mean centered on f(x)'s value $p(y) = N(y \mid f(x))$
- Now, linear regression is:

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

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





$p(y|x) = N(u, \sigma^2) \Rightarrow f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp(-\frac{(x-u)^2}{2\sigma^2}) = pdf$

$N(0,1) \Rightarrow f(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right) = pdf$

$N(u,1) \Rightarrow pdf = p(y|u) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(y-u)^2}{2}\right)$; when $u = f(x) = \theta^T x + b$

Tony Jebara, Columbia University

Probability Models

$p(y|f(x)) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta^T x - b)^2}{2}\right)$

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

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

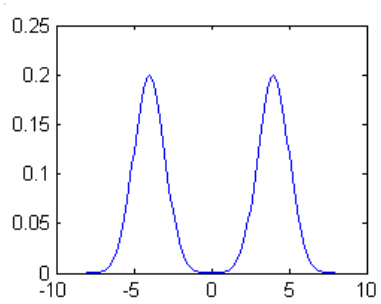
$$= -N \log\left(\sqrt{2\pi}\right) - \sum_{i=1}^{N} \frac{1}{2}\left(y_i - f(x_i)\right)^2$$

pdf: $N(u,1)$

$L(\theta) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(y-f(x))^2}{2}\right)$ / $\log L(\theta) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(y-f(x))^2}{2}\right)$

$= \frac{-N}{2}\log 2\pi + \sum_{i=1}^{N} -\frac{(y-f(\theta))^2}{2} = -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\frac{(y-f(\theta))^2}{2}$

Probability Models



- Can extend probability model to 2 bumps:

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

- Each mean can be a linear regression fn.

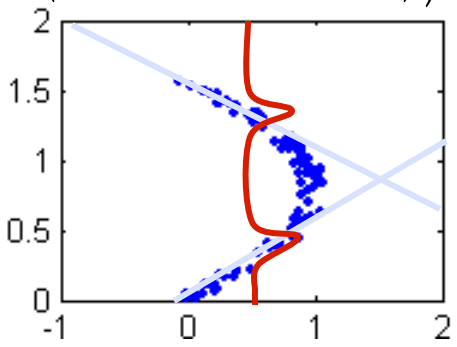
$$p(y \mid x, \Theta) = \tfrac{1}{2} N(y \mid f_1(x)) + \tfrac{1}{2} N(y \mid f_2(x))$$

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

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

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

- Maximize l(θ) 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):

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

$$\prod_{i=1}^{N} p(y|x) = \prod_{i=1}^{N} f(x)^{y_i}(1-f(x))^{1-y_i}$$

Probability Models

- Now linear classification is:

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

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

$$\sum_{i=1}^{N} \log p(y_i \mid x_i) = \sum_{i=1}^{N} \log f(x_i)^{y_i} \left(1 - f(x_i)\right)^{1-y_i}$$

$$= \sum_{i=1}^{N} y_i \log f(x_i) + (1 - y_i) \log\left(1 - f(x_i)\right)$$

$$= \sum_{i \in class1} \log f(x_i) + \sum_{i \in class0} \log\left(1 - f(x_i)\right)$$

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

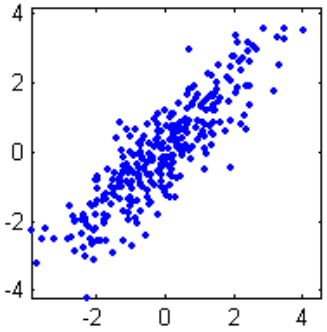
$$f(x) = sigmoid\left(\theta^T x + b\right) \in [0,1]$$

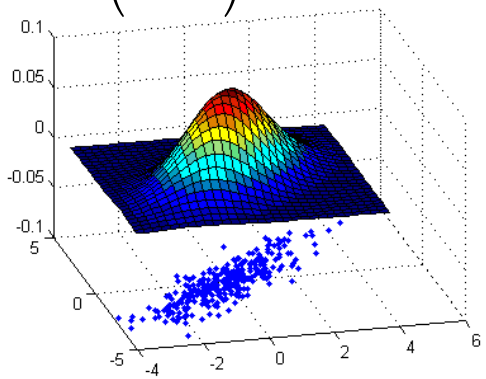
- Called logistic regression → *new loss function*
- Do gradient descent, similar to logistic output neural net!
- Can also handle multi-layer f(x) and do backprop again!

Generative Probability Models

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

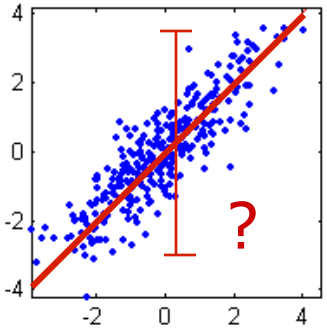
E.g. *describe* data with Multi-Dim. Gaussian (later…)





- 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



PDF = probability Density Function.

Properties of PDFs

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

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

- 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:

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

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…

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

IMPORTANT

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

posterior likelihood

$$p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta)\, p(\theta)}{p(\mathcal{X})}$$

likelihood → $p(\mathcal{X} \mid \theta)$; **prior** → $p(\theta)$; **posterior** → $p(\theta \mid \mathcal{X})$; **evidence** → $p(\mathcal{X})$

Properties of PDFs

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

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

- Properties:

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

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

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

- Mean: expected value for x

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

example: speeding ticket

Fine=0$	Fine=20$
0.8	0.2

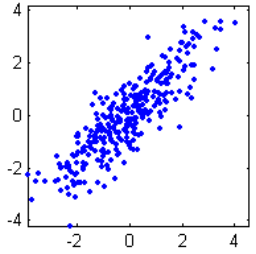
expected cost of speeding?
f(x=0)=0, f(x=1)=20
p(x=0)=0.8, p(x=1)=0.2

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

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

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

Properties of PDFs



- Covariance: how strongly x and y vary together

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

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

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

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

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

- Sample Mean: $E\{x\} \simeq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Sample Var: $E\{(x - E(x))^2\} \simeq \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$
- Sample Cov: $E\{(x - E(x))(y - E(y))\} \simeq \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$

More Properties of PDFs

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

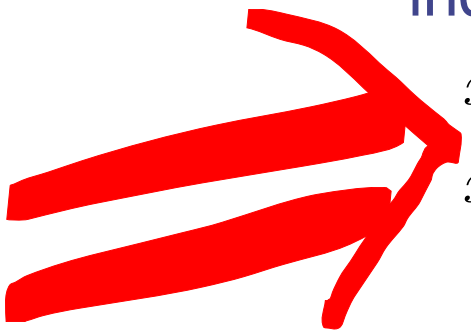
$$x \perp\!\!\!\perp y \quad \rightarrow \quad p(x,y) = p(x)p(y)$$
$$x \perp\!\!\!\perp y \quad \rightarrow \quad p(x \mid y) = p(x)$$

also note in this case:

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

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

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



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:

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

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

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

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

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

Θ: model
X: data

The IID Assumption

- Bayes rule says likelihood is probability of data given model

likelihood

posterior $p(\theta \mid \mathcal{X}) = \dfrac{p(\mathcal{X} \mid \theta)\, p(\theta)}{p(\mathcal{X})}$ prior

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:

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

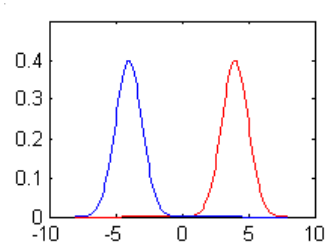
- Learn conditional $p(y \mid x, \Theta)$ by max conditional likelihood:

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

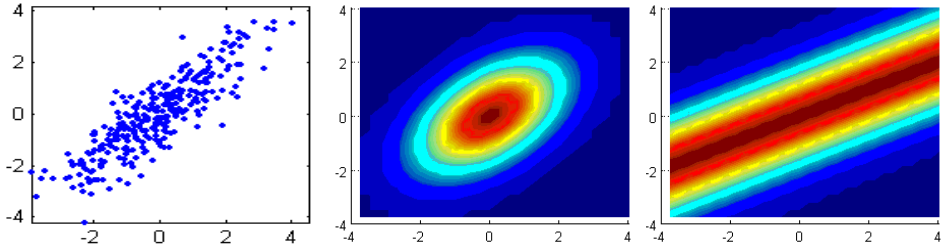
Uses of PDFs

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

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



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



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

$$\hat{y} = \begin{cases} \arg\max_y p(y \mid x) \\ E_{p(y|x)}\{y\} \end{cases}$$

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

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