Machine Learning:
The Probabilistic Perspective

Konstantin Tretyakov

http://kt.era.ee

AACIMP 2012
August 2012

STACC Software Technology and Applications Competence Center

BIIT

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS 1632

So far…

- Machine learning is important and interesting
- The general concept:

Fitting models to data

So far…

- Machine learning is important and interesting
- The general concept:

Fitting models to data

Optimization

Probability Theory

So far…

- Instance-based methods
- Tree learning methods
- The "soul" of machine learning:

$$\boxed{\operatorname{argmin}_{\boldsymbol{w}} \mathrm{Error}(\mathrm{Data}, \boldsymbol{w}) + \lambda\, \mathrm{Complexity}(\boldsymbol{w})}$$

- Particular models:
  - OLS regression (ℓ_2-loss, 0-penalty regression)
  - Ridge regression (ℓ_2-loss, ℓ_2-penalty regression)

So far…

- Analytic vs iterative optimization
- Batch vs on-line optimization

- Training / Test sets, cross-validation

Today

Why should the model, tuned on the **training set**, **generalize** to the test set?

The "No Free Lunch" Principle

Learning purely from data is, in general, impossible

X	Y	Output
0	0	False
0	1	True
1	0	True
1	1	?

The "No Free Lunch" Principle

Learning purely from data is, in general, impossible

- Is it good or bad?
  -
- What should we do to enable learning?
  -

The “No Free Lunch” Principle

Learning purely from data is, in general, impossible

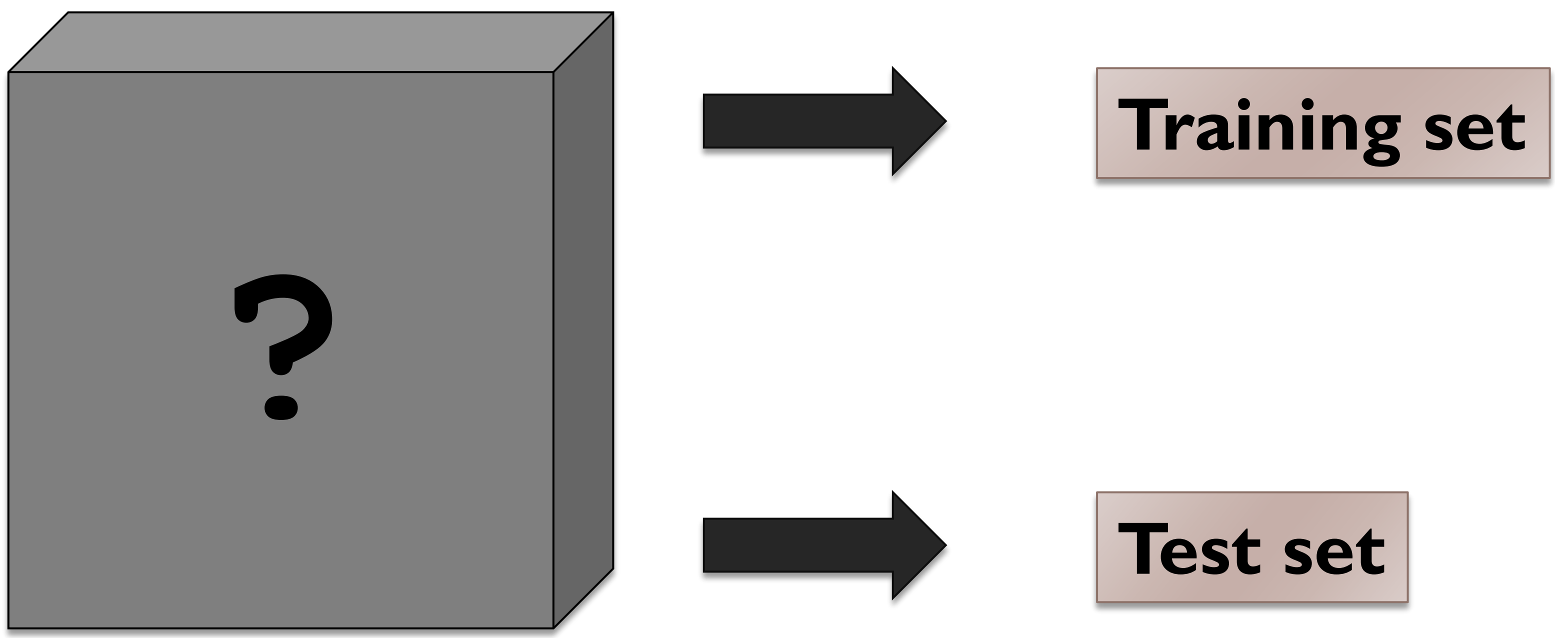
- Is it good or bad?
  - Good for cryptographers, bad for data miners
- What should we do to enable learning?
  - Introduce **assumptions about data** (“inductive bias”):
    1. **How does existing data relate to the future data?**
    2. **What is the system we are learning?**

The "No Free Lunch" Principle

Learning purely from data is, in general, impossible

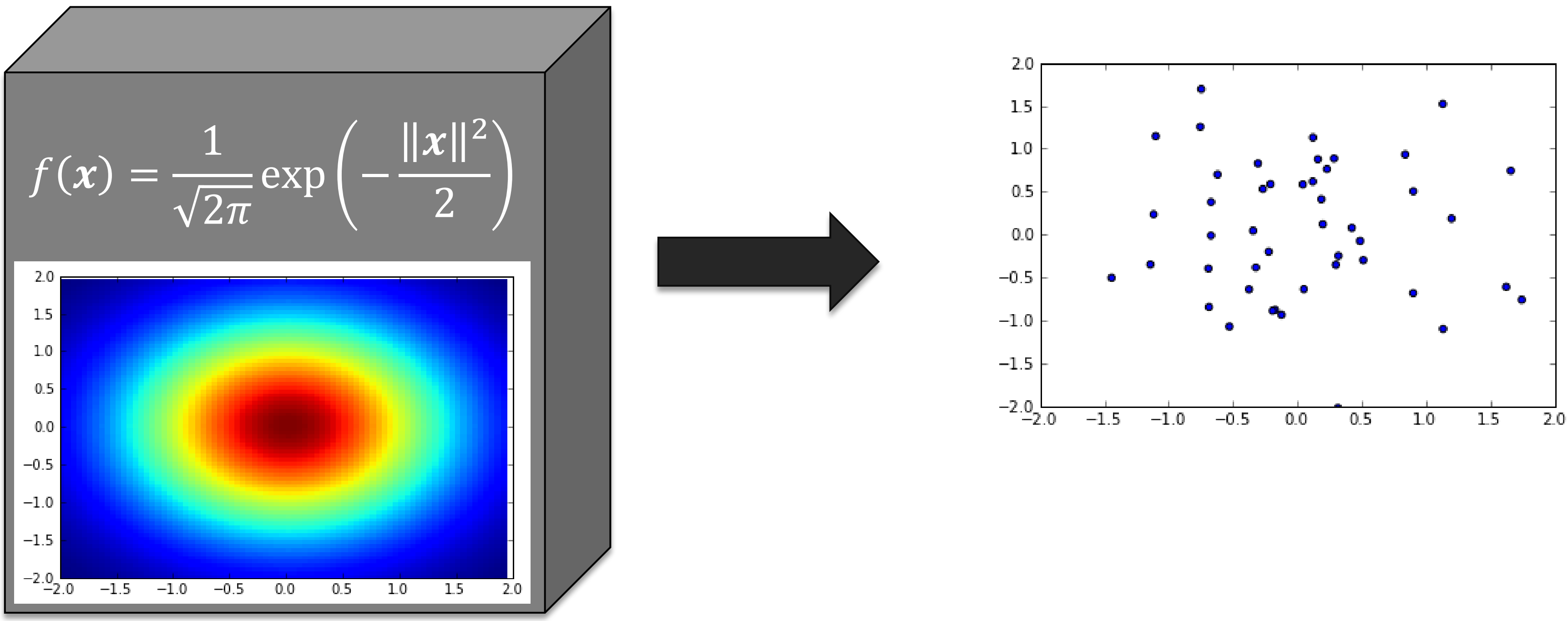
- Is it good or bad?
  - Good for cryptographers, bad for data miners
- What should we do to enable learning?
  - Introduce **assumptions about data** ("inductive bias"):
    1. **How does existing data relate to the future data?**
    2. **What is the system we are learning?**

How does existing data relate to future data?



X	P(X)
heads	0.5
tails	0.5

heads,
heads,
tails,
heads,
tails,
…


$f(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)$

Probability theory

$\mathrm{B}(n, p)$

$\mathrm{Hg}(m, n, M, N)$

$\mathrm{P}(\lambda)$

$N(\mu, \Sigma)$

$\mathrm{Zipf}(\alpha)$

$\mathrm{Exp}(\lambda)$

$\beta(a, b)$

$\mathrm{Be}(p)$

$\mathrm{U}(a, b)$

…

P

V · T · E

Probability distributions [hide]

Discrete univariate with finite support [hide]

Benford · Bernoulli · Beta-binomial · binomial · categorical · hypergeometric · Poisson binomial · Rademacher · discrete uniform · Zipf · Zipf-Mandelbrot

Discrete univariate with infinite support [hide]

beta negative binomial · Boltzmann · Conway–Maxwell–Poisson · discrete phase-type · Delaporte · extended negative binomial · Gauss–Kuzmin · geometric · logarithmic · negative binomial · parabolic fractal · Poisson · Skellam · Yule–Simon · zeta

Continuous univariate supported on a bounded interval, e.g. [0,1] [hide]

Arcsine · ARGUS · Balding-Nichols · Bates · Beta · Beta rectangular · Irwin–Hall · Kumaraswamy · logit-normal · Noncentral beta · raised cosine · triangular · U-quadratic · uniform · Wigner semicircle

Continuous univariate supported on a semi-infinite interval, usually [0,∞) [hide]

Benini · Benktander 1st kind · Benktander 2nd kind · Beta prime · Bose–Einstein · Burr · chi-squared · chi · Coxian · Dagum · Davis · Erlang · exponential · *F* · Fermi–Dirac · folded normal · Fréchet · Gamma · generalized inverse Gaussian · half-logistic · half-normal · Hotelling's T-squared · hyper-exponential · hypoexponential · inverse chi-squared (scaled-inverse-chi-squared) · inverse Gaussian · inverse gamma · Kolmogorov · Lévy · log-Cauchy · log-Laplace · log-logistic · log-normal · Maxwell–Boltzmann · Maxwell speed · Mittag–Leffler · Nakagami · noncentral chi-squared · Pareto · phase-type · Rayleigh · relativistic Breit–Wigner · Rice · Rosin–Rammler · shifted Gompertz · truncated normal · type-2 Gumbel · Weibull · Wilks' lambda

Continuous univariate supported on the whole real line (−∞, ∞) [hide]

Cauchy · exponential power · Fisher's z · generalized normal · generalized hyperbolic · geometric stable · Gumbel · Holtsmark · hyperbolic secant · Landau · Laplace · Linnik · logistic · noncentral t · normal (Gaussian) · normal-inverse Gaussian · skew normal · slash · stable · Student's *t* · type-1 Gumbel · variance-gamma · Voigt

Continuous univariate with support whose type varies [hide]

generalized extreme value · generalized Pareto · Tukey lambda · q-Gaussian · q-exponential · shifted log-logistic

Mixed continuous-discrete univariate distributions [hide]

rectified Gaussian

Multivariate (joint) [hide]

Discrete: Ewens · multinomial · Dirichlet-multinomial · negative multinomial

Continuous: Dirichlet · Generalized Dirichlet · multivariate normal · Multivariate stable · multivariate Student · normal-scaled inverse gamma · normal-gamma

Matrix-valued: inverse matrix gamma · inverse-Wishart · matrix normal · matrix t · matrix gamma · normal-inverse-Wishart · normal-Wishart · Wishart

Directional [hide]

Univariate (circular) directional: Circular uniform · univariate von Mises · wrapped normal · wrapped Cauchy · wrapped exponential · wrapped Lévy

Bivariate (spherical): Kent · *Bivariate (toroidal)*: bivariate von Mises

Multivariate: von Mises–Fisher · Bingham

Degenerate and singular [hide]

Degenerate: discrete degenerate · Dirac delta function

Singular: Cantor

Families [hide]

Probability theory

$$\begin{pmatrix} 10 + \sin(N(0,2) \cdot B(0.1)) \\ F(U(0,1)) \end{pmatrix}$$

Probability theory

```
from numpy.random import beta, binomial,
chisquare, dirichlet, exponential, f, gamma,
geometric, gumbel, hypergeometric, ...
```

```
>>> numpy.random.seed(1)
>>> binomial(10, 0.2)
::: 2
```

Probability theory

```
from scipy.stats.distributions import beta,
binom, chisquare, ...

>>> numpy.random.seed(1)
>>> X = binom(10, 0.2)
>>> X.rvs()
::: 2

>>> X.pmf(2), X.cdf(2), X.mean(), X.std(), ...
```

Everything is Probabilistic?

What is your height?

Everything is Probabilistic?

What is your height?

Is it a fixed number?

Everything is Probabilistic?

What is your height?

Is it a fixed number?

- Frequentist: **Yes, it is**, we just don't know it precisely.
- Bayesian: **No, it is not**. The result is a **distribution**.

Everything is Probabilistic?

What is your height?

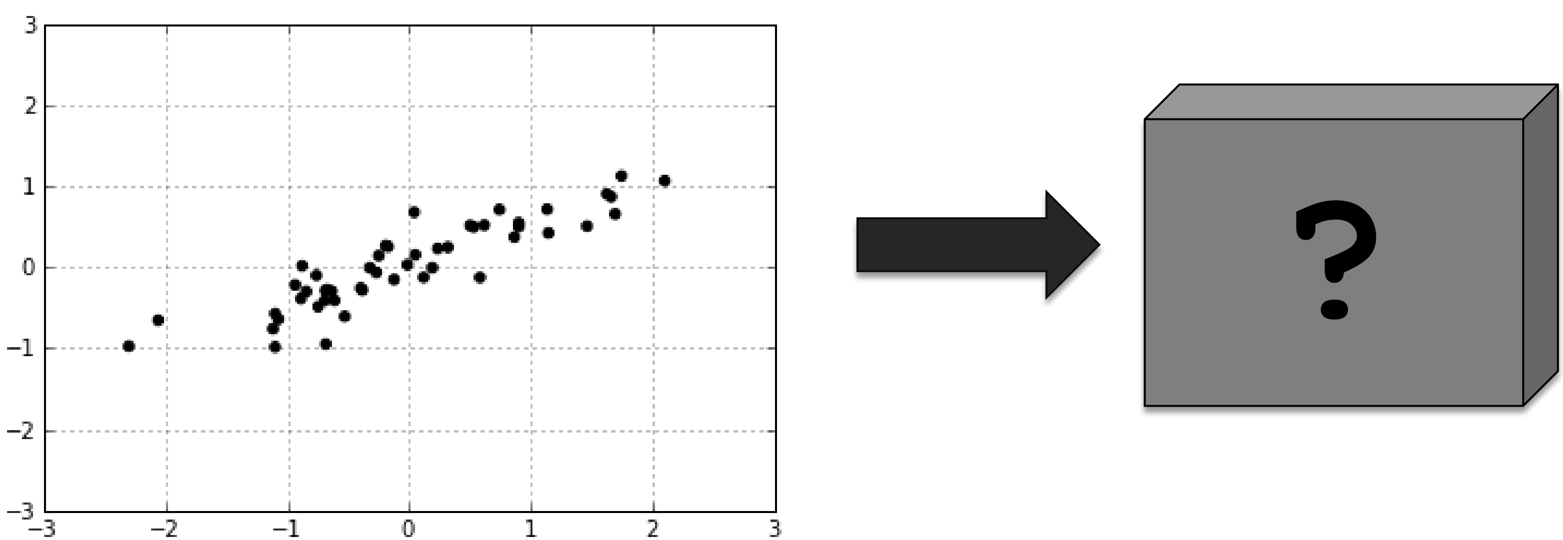
Is it a fixed number?

- Frequentist: **Yes, it is**, we just don't know it precisely.
- Bayesian: **No, it is not**. The result is a **distribution**.

In any case, we need probabilistic reasoning.

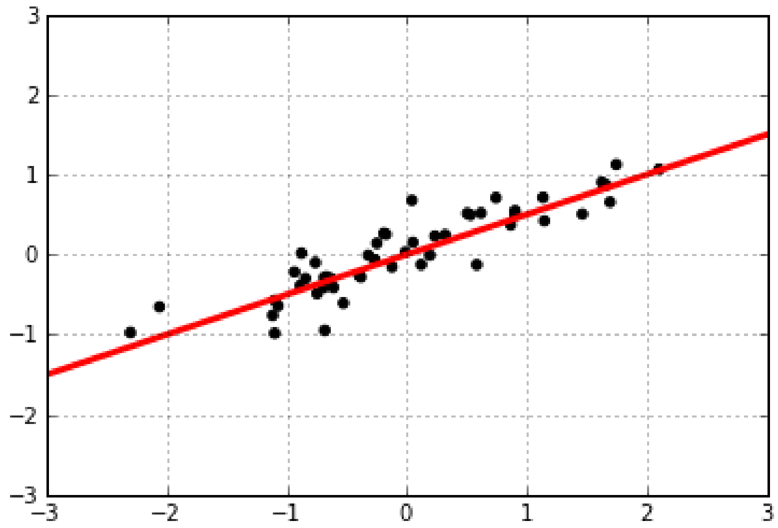
Statistics & Decision Theory

- Statistics
  - How do we **infer** a probabilistic **model** based on data?



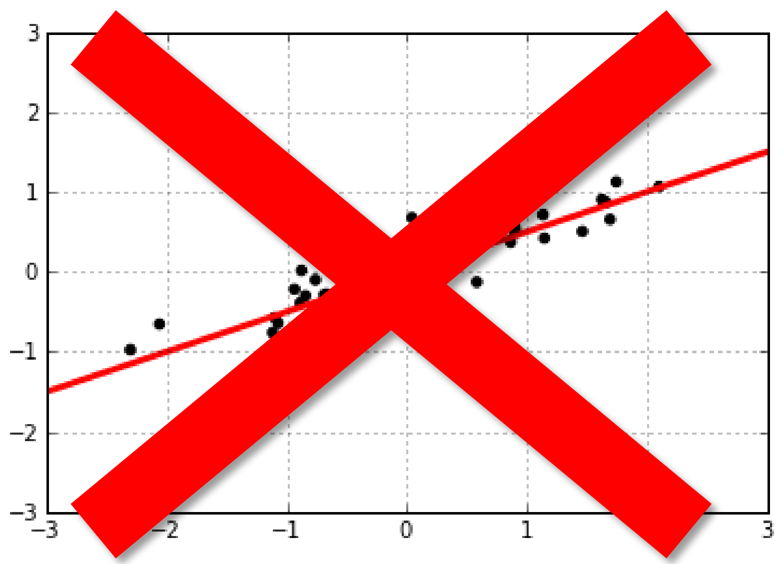
Statistics & Decision Theory

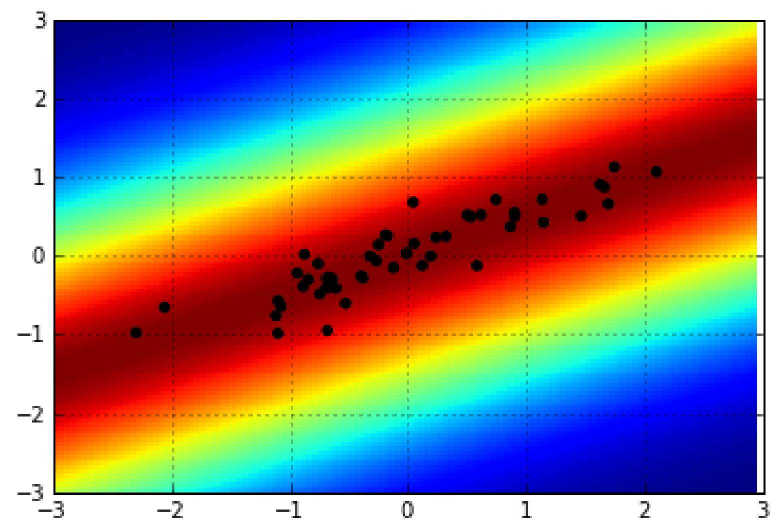
- Statistics
  - How do we **infer** a probabilistic **model** based on data?



Statistics & Decision Theory

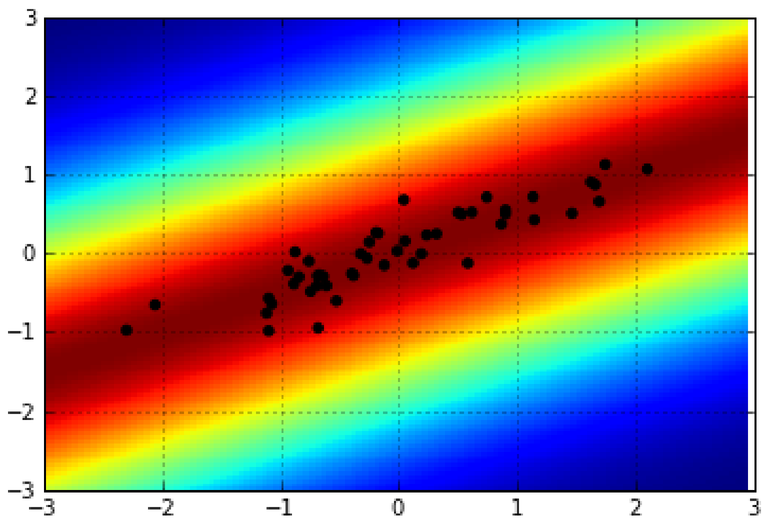
- Statistics
  - How do we **infer** a probabilistic **model** based on data?





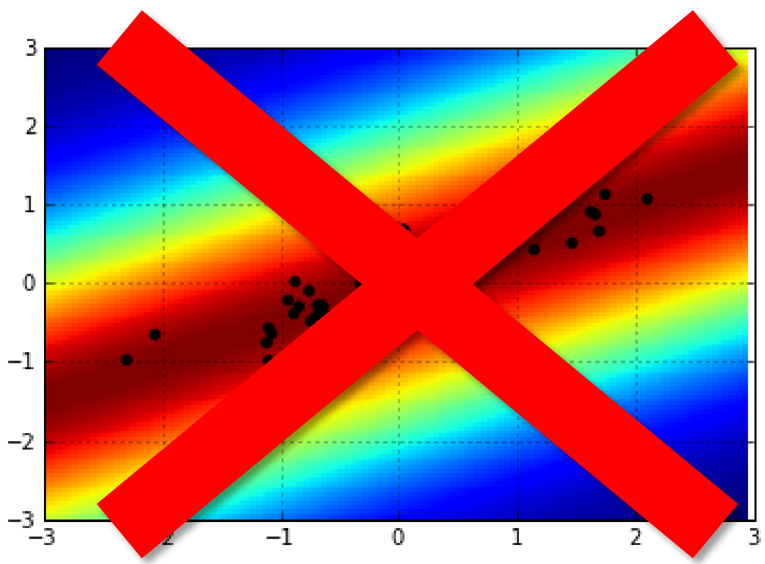
Statistics & Decision Theory

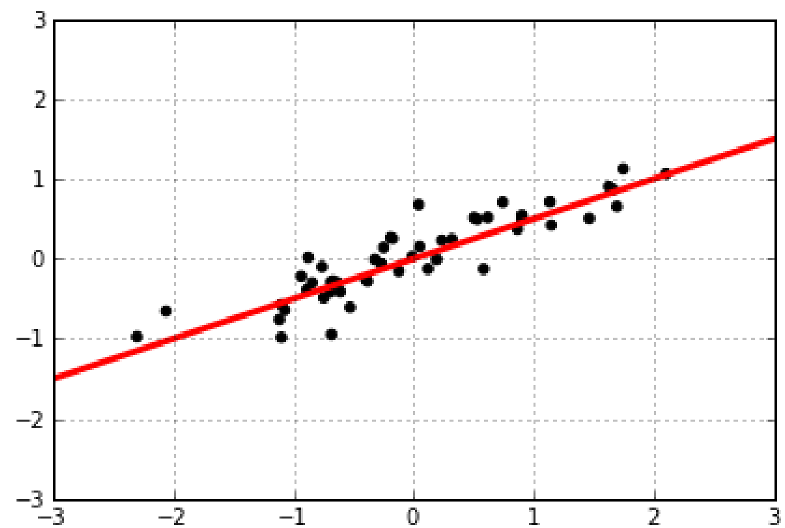
- Decision theory
  - How do we **use** a probabilistic model **to predict**?



Statistics & Decision Theory

- Decision theory
  - How do we **use** a probabilistic model **to predict**?

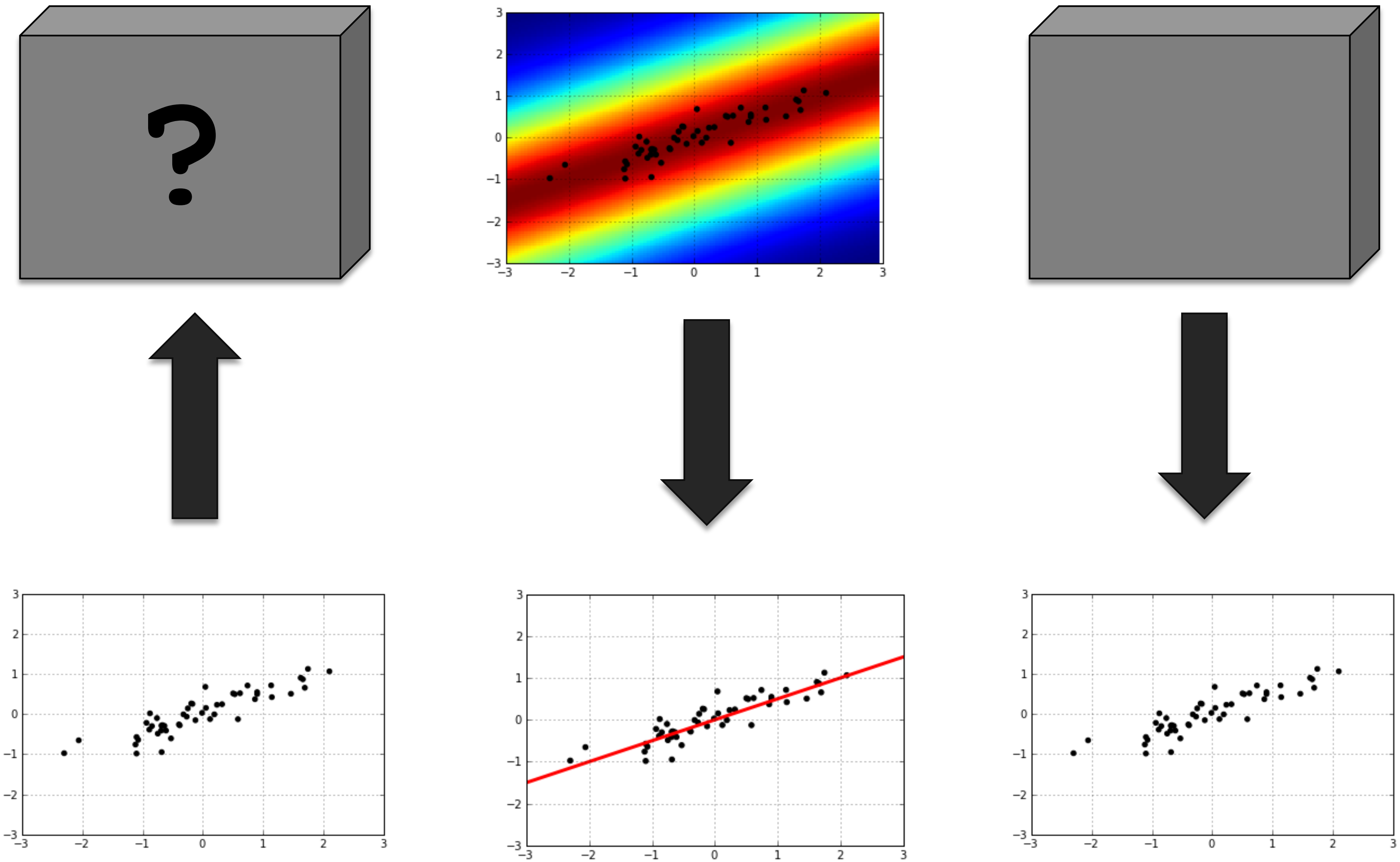




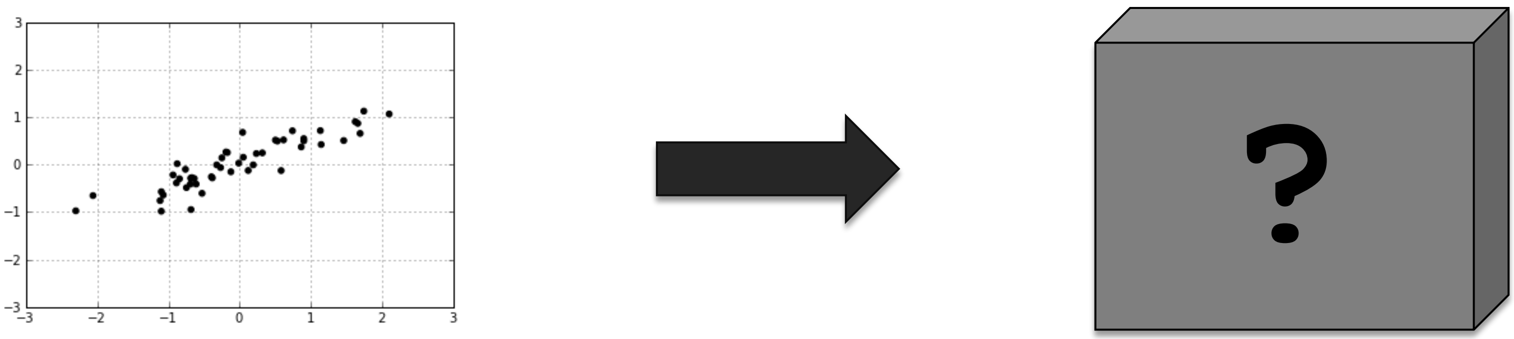
Quiz

- Model, trained on the training set might work well on the test set because:
  - Because we **assume** a single underlying mechanism.
  - Because we **use statistical inference** to infer the mechanism.
  - Because we **use decision theory** to produce optimal decisions.

Quiz

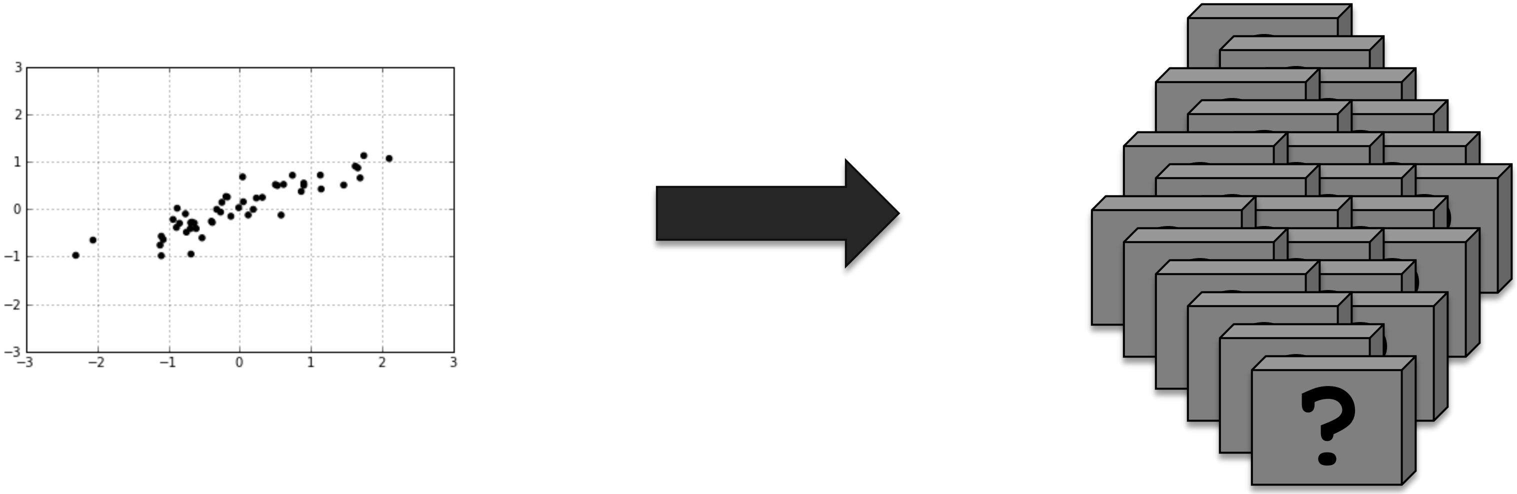


Statistics

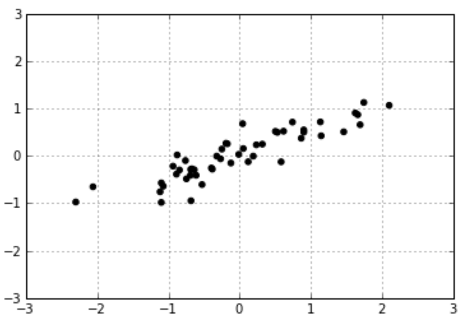


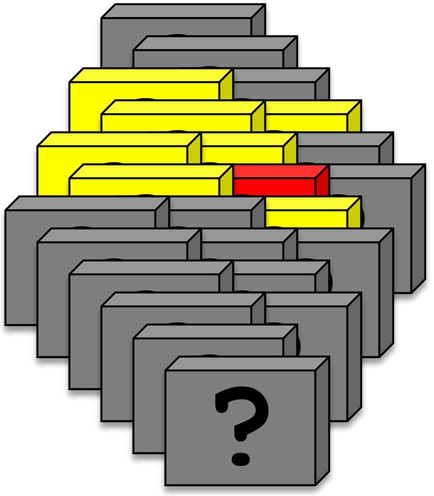
Statistics

Space of candidate models



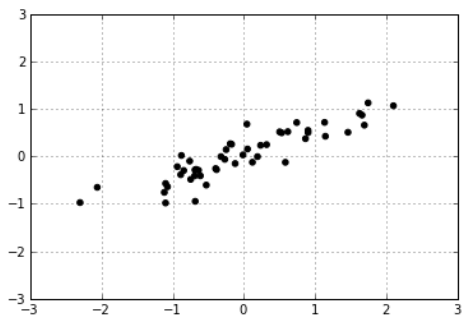
Statistics

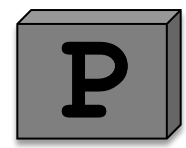




Statistics

Hypothesis testing

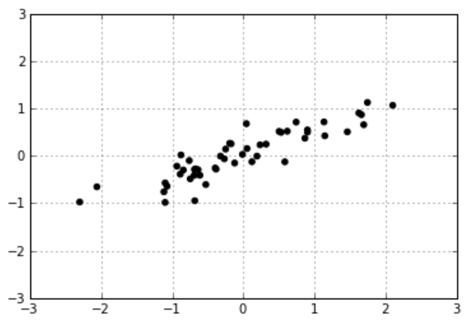


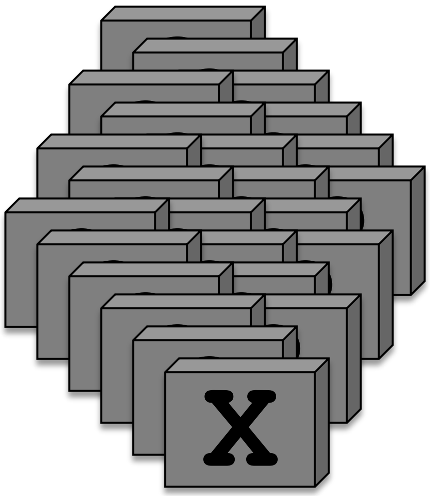


or not?

Statistics

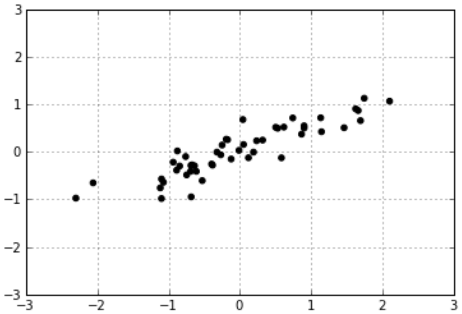
Model selection





Statistics

Parameter inference



$$P(X \mid \theta)$$

Parameter inference

Biased coin

$\mathrm{Be}(p)$

X	P(X)
1	p
0	1-p

1,1,0,1,1

$$n_1 = 4$$
$$n_0 = 1$$

Maximum Likelihood Estimation

- Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- Example:
  - Model: Be(0.5)
  - Data: 1,1,0,1,1
  - Likelihood: ?

Maximum Likelihood Estimation

- Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- Example:
  - Model: Be(0.5)
  - Data: $1,1,0,1,1$
  - Likelihood: $0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 2^{-5}$

0.03125

Maximum Likelihood Estimation

- Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- Example:
  - Model: Be(0.2)
  - Data: 1,1,0,1,1
  - Likelihood: ?

Maximum Likelihood Estimation

- Data Likelihood:

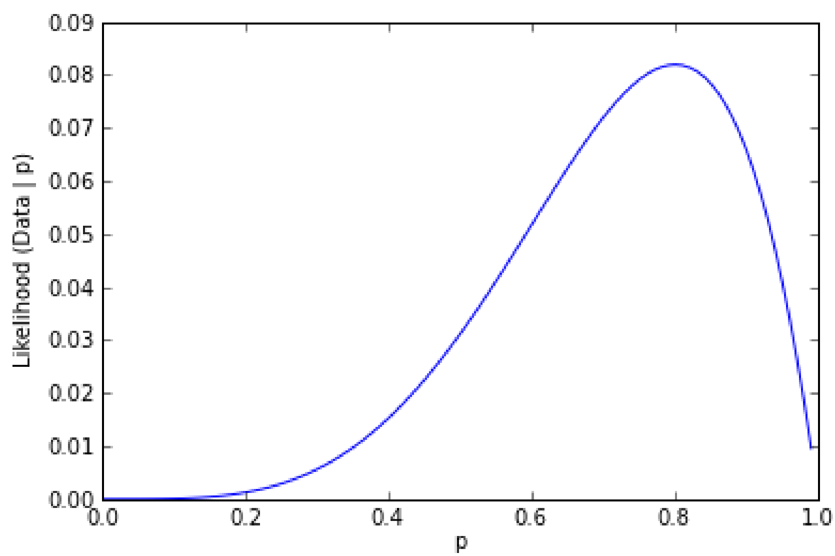
$$\Pr[\text{Data} \mid \text{Model}]$$

- Example:
  - Model: Be(0.2)
  - Data: 1,1,0,1,1
  - Likelihood: $0.2 \cdot 0.2 \cdot 0.8 \cdot 0.2 \cdot 0.2 = 0.2^4 \cdot 0.8$

0.00128

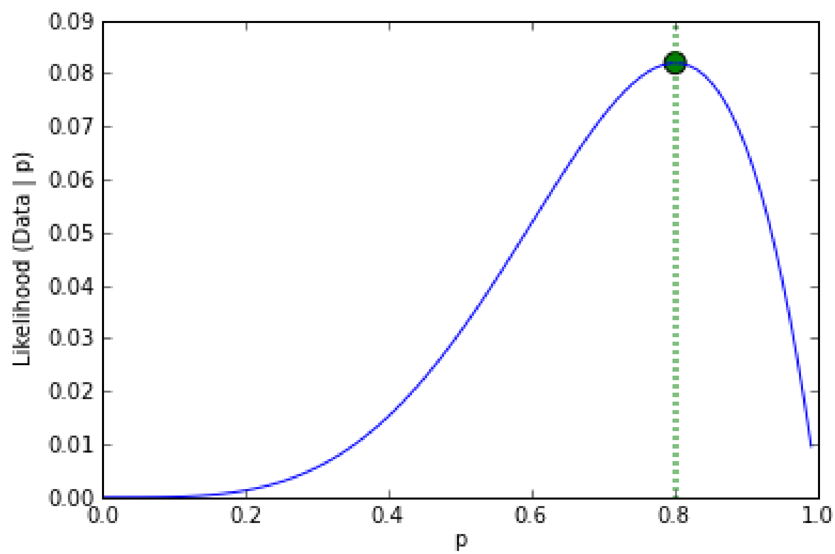
Maximum Likelihood Estimation

- Example:
  - Model: Be(p)
  - Data: 1,1,0,1,1
  - Likelihood: $p \cdot p \cdot (1-p) \cdot p \cdot p = p^{n_1}(1-p)^{n_0}$



Maximum Likelihood Estimation

- Example:
  - Model: Be(p)
  - Data: 1,1,0,1,1
  - Likelihood: $p \cdot p \cdot (1-p) \cdot p \cdot p = p^{n_1}(1-p)^{n_0}$



$$\hat{p} = \frac{n_1}{n_0 + n_1}$$

Maximum Likelihood Estimation

- Maximum Likelihood Estimation:

$$\operatorname{argmax}_{\mathrm{Model}} \Pr(\mathrm{Data} \mid \mathrm{Model})$$

Problems of MLE

- You are on a trip in an exotic country and you meet a person who happens to be from Ukraine.
- Is he a member of the Rada?

Problems of MLE

- Data: “X is from Ukraine”
- Models:
  - “X is a member of Rada”,
  - “X is not a member of Rada”

Problems of MLE

- Data: “X is from Ukraine”
- Models:
  - “X is a member of Rada”,
  - “X is not a member of Rada”
- Likelihoods:
  - P(X is from Ukraine | X is a member of Rada) =
  - P(X is from Ukraine | X is not a member of Rada) =

Problems of MLE

- Data: "X is from Ukraine"
- Models:
  - "X is a member of Rada",
  - "X is not a member of Rada"
- Likelihoods:
  - P(X is from Ukraine | X is a member of Rada) = 1
  - P(X is from Ukraine | X is not a member of Rada) = $\frac{45}{7000}$

Problems of MLE

- Data: "X is from Ukraine"
- Models:
  - "X is a member of Rada",
  -
- L
  - P(X is from Ukraine | X is a member of Rada) = 1
  - P(X is from Ukraine | X is not a member of Rada) = $\frac{45}{7000}$

MLE treats all candidate models as equal and can thus overfit

Maximum A-posteriori Estimation

- Maximum Likelihood Estimate (MLE):

$$\operatorname{argmax}_{\text{Model}} \ \Pr(\text{Data}\,|\,\text{Model})$$

- Maximum A-posteriori Estimate (MAP):

$$\operatorname{argmax}_{\text{Model}} \ \Pr(\text{Model}\,|\,\text{Data})$$

MAP Estimation

$$\operatorname{argmax}_{\mathrm{Model}} \ \Pr(\mathrm{Model}|\mathrm{Data})$$

MAP Estimation

$$\text{argmax}_{\text{Model}} \ \Pr(\text{Model}|\text{Data})$$

$$\text{argmax}_{\text{Model}} \frac{\Pr(\text{Model}, \ \text{Data})}{\Pr(\text{Data})}$$

$$\text{argmax}_{\text{Model}} \ \Pr(\text{Model}, \text{Data})$$

MAP Estimation

$$\mathrm{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Model}|\mathrm{Data})$$

$$\mathrm{argmax}_{\mathrm{Model}} \frac{\Pr(\mathrm{Model},\ \mathrm{Data})}{\Pr(\mathrm{Data})}$$

$$\mathrm{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Model}, \mathrm{Data})$$

$$\mathrm{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})$$

MAP Estimation

$$\mathrm{argmax}_{\mathrm{Model}}\ \boxed{\Pr(\mathrm{Model}|\mathrm{Data})}$$

Model posterior

$$\mathrm{argmax}_{\mathrm{Model}} \frac{\Pr(\mathrm{Model},\mathrm{Data})}{\Pr(\mathrm{Data})}$$

$$\mathrm{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Model},\mathrm{Data})$$

$$\mathrm{argmax}_{\mathrm{Model}}\ \boxed{\Pr(\mathrm{Data} \mid \mathrm{Model})} \cdot \boxed{\Pr(\mathrm{Model})}$$

Likelihood

Model prior

Summary

- Maximum Likelihood Estimate (MLE):

$$\text{argmax}_{\text{Model}} \ \Pr(\text{Data} \mid \text{Model})$$

- Maximum A-posteriori Estimate (MAP):

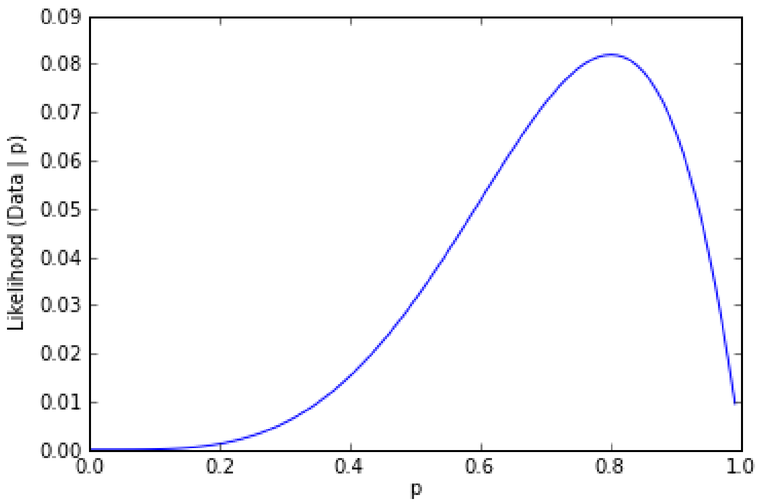
$$\text{argmax}_{\text{Model}} \ \Pr(\text{Data} \mid \text{Model}) \ \Pr(\text{Model})$$

MAP Estimation

- Model: Be(p)

Data: 1,1,0,1,1

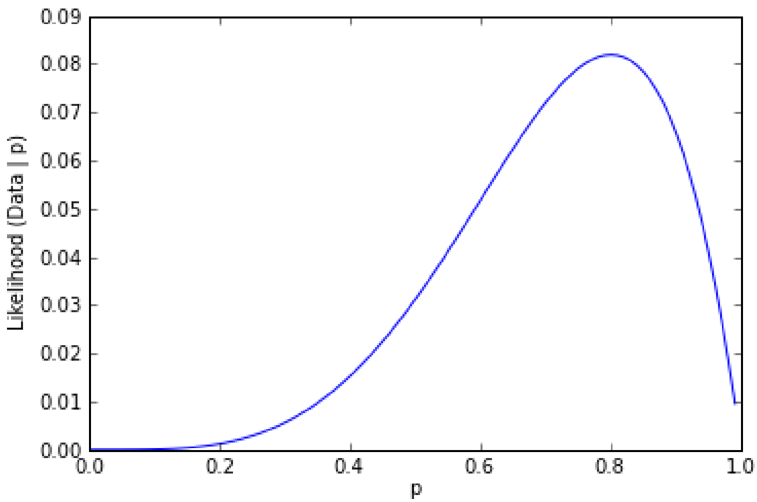
Likelihood: $p^4(1-p)$

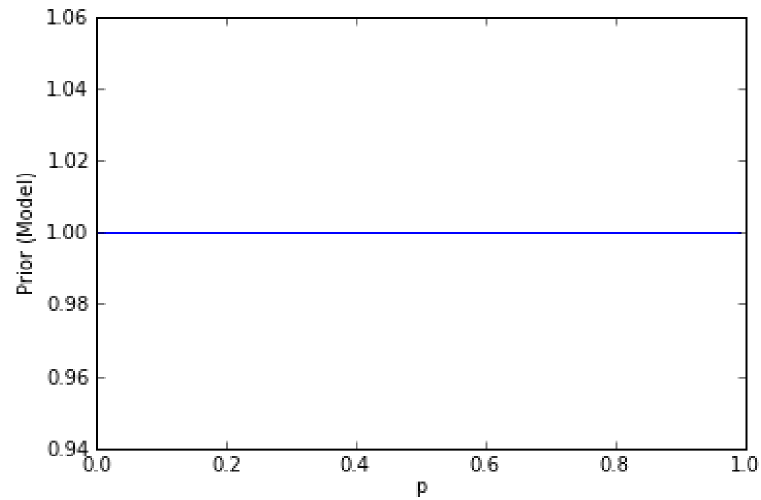


MAP Estimation

- Model: Be(p) Data: 1,1,0,1,1

Likelihood: $p^4(1-p)$ Prior: $U(0,1)$





$$\hat{p}_{MAP} = \hat{p}_{MLE} = \frac{n_1}{n_0 + n_1}$$

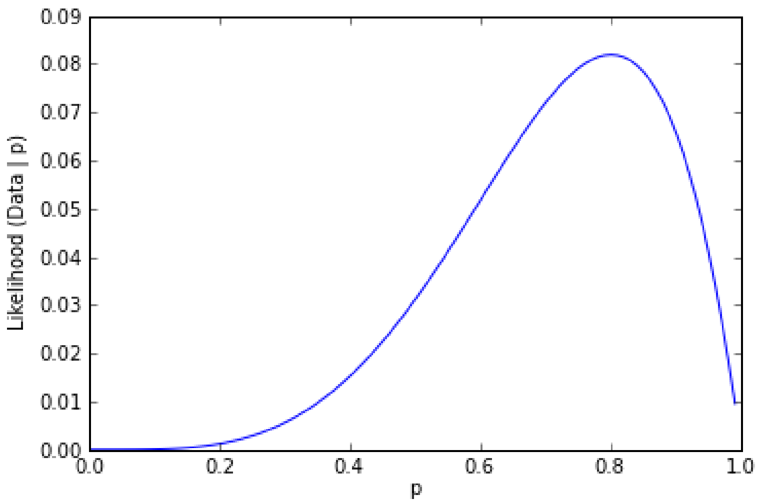
MAP Estimation

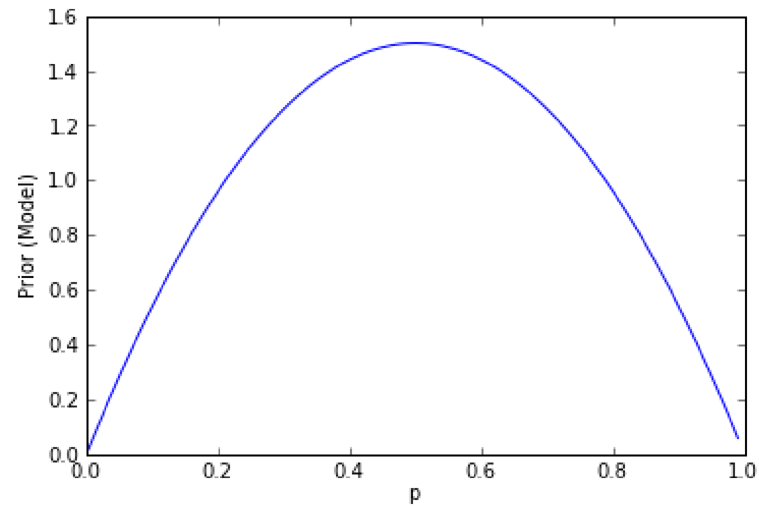
- Model: Be(p)

Data: 1,1,0,1,1

Likelihood: $p^4(1-p)$

Prior: $\mathrm{Beta}(2,2)$

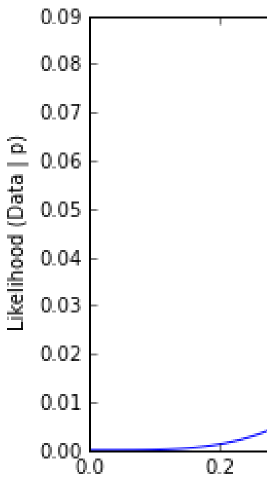


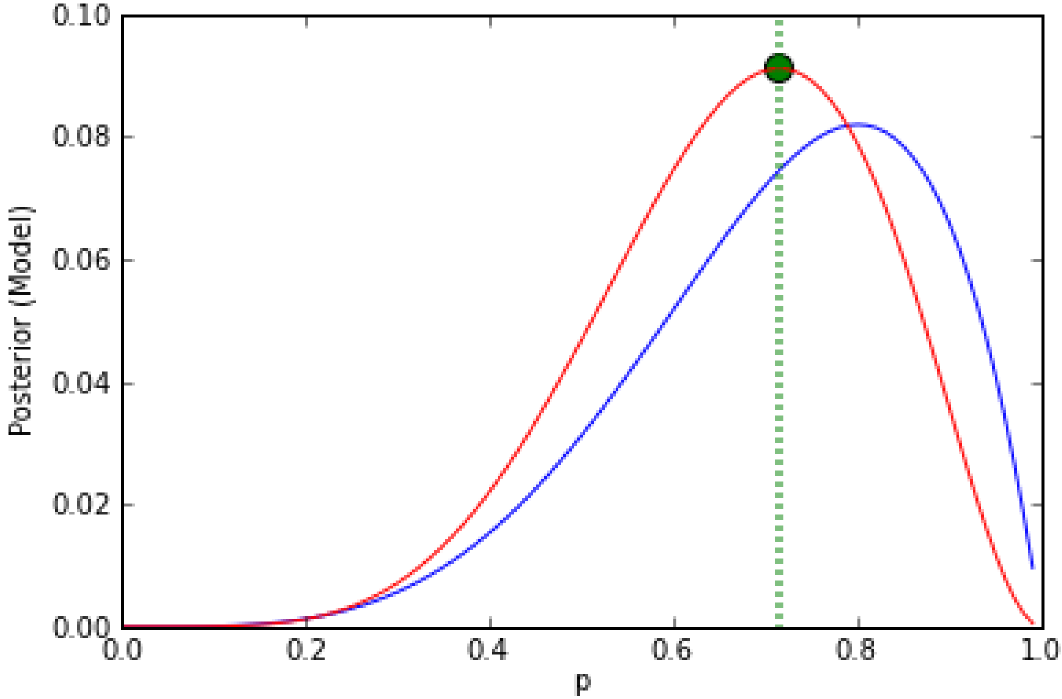


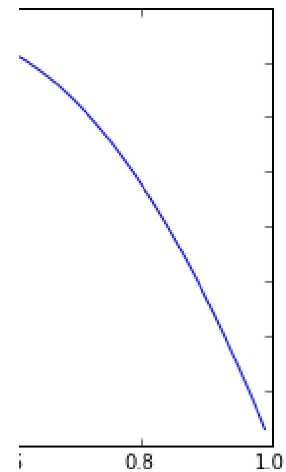
MAP Estimation

- Model: Be(p) Data: 1,1,0,1,1

  Likelih $2, 2)$







$$\hat{p}_{MAP} = \frac{n_1 + 1}{n_0 + n_1 + 2}$$

MAP Estimation

$$\operatorname{argmax}_{\mathrm{Model}} \Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})$$

MAP Estimation

$$\text{argmax}_{\text{Model}}\ \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

$$\text{argmax}_{\text{Model}}\ \log\left(\Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})\right)$$

MAP Estimation

$$\mathrm{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})$$

$$\mathrm{argmax}_{\mathrm{Model}}\ \log\left(\Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})\right)$$

$$\mathrm{argmax}_{\mathrm{Model}} \log \Pr(\mathrm{Data}|\mathrm{Model}) + \log \Pr(\mathrm{Model})$$

MAP Estimation

$$\operatorname{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})$$

$$\operatorname{argmax}_{\mathrm{Model}}\ \log\left(\Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})\right)$$

$$\operatorname{argmax}_{\mathrm{Model}} \boxed{\log \Pr(\mathrm{Data}|\mathrm{Model}) + \log \Pr(\mathrm{Model})}$$

MAP Estimation

$$\operatorname{argmax}_{\mathrm{Model}}\ \Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})$$

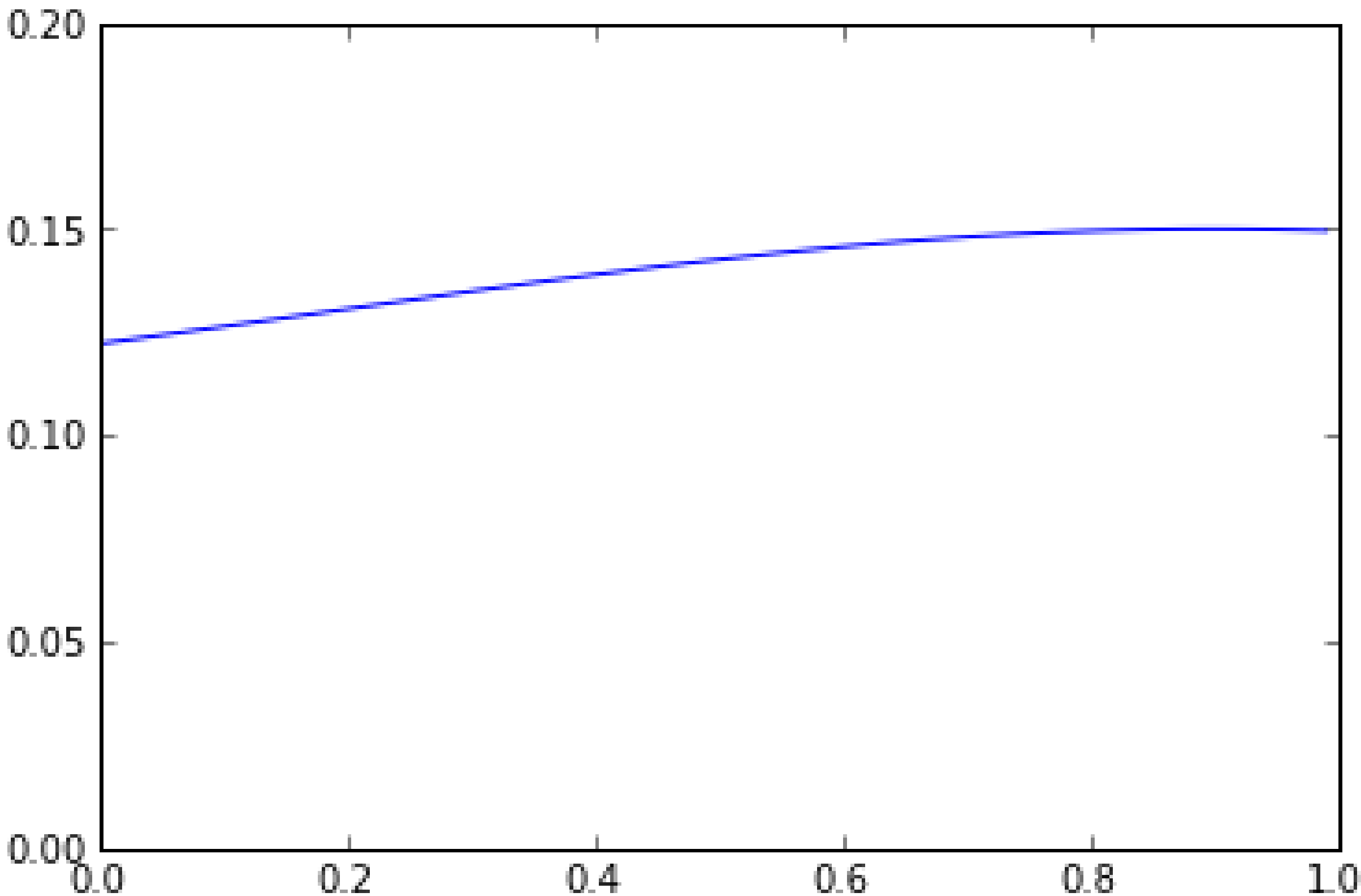
$$\operatorname{argmax}_{\mathrm{Model}}\ \log\left(\Pr(\mathrm{Data} \mid \mathrm{Model}) \cdot \Pr(\mathrm{Model})\right)$$

$$\operatorname{argmax}_{\mathrm{Model}} \boxed{\log \Pr(\mathrm{Data}|\mathrm{Model}) + \log \Pr(\mathrm{Model})}$$

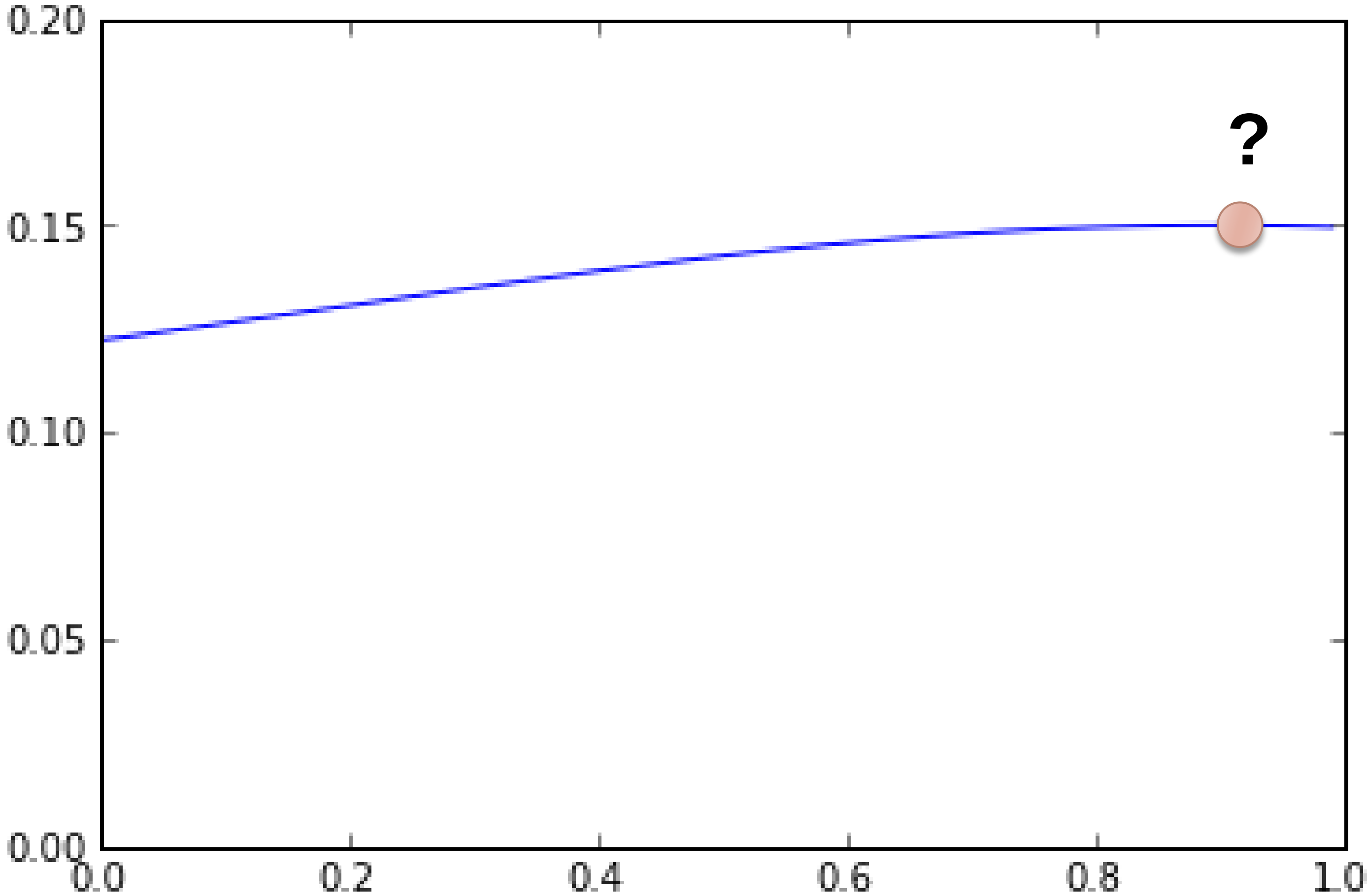
$$\operatorname{argmin}_{\boldsymbol{w}}\ \mathrm{Error}(\mathrm{Data}, \boldsymbol{w}) + \mathrm{Complexity}(\boldsymbol{w})$$

Problems of MAP estimation

Problems of MAP estimation



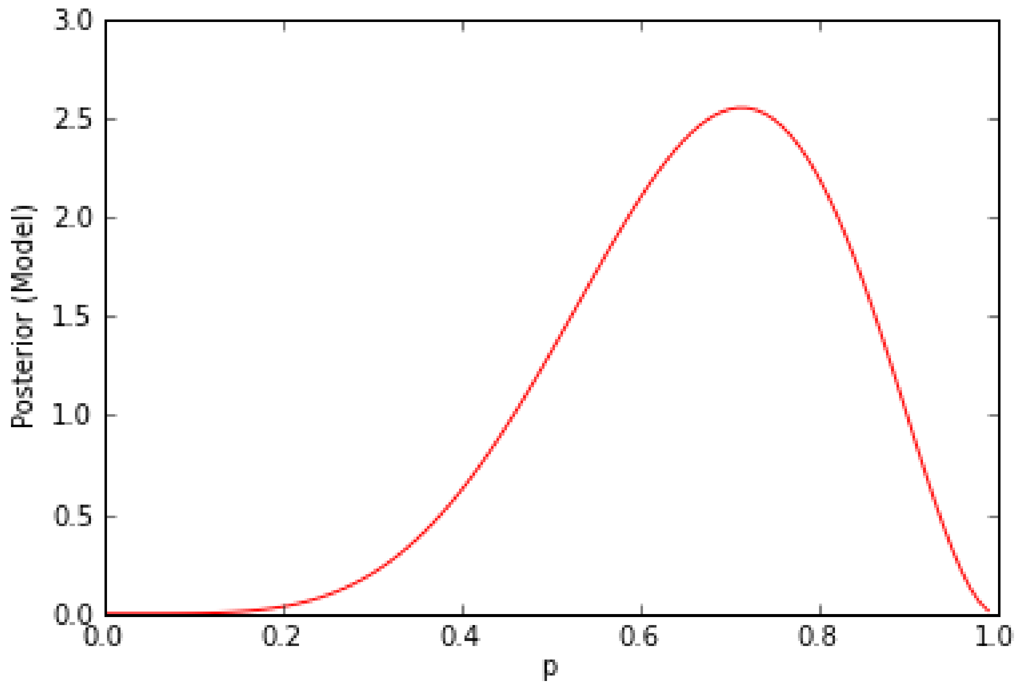
Problems of MAP estimation



Bayesian estimation

- Pick the model with minimal *expected risk*

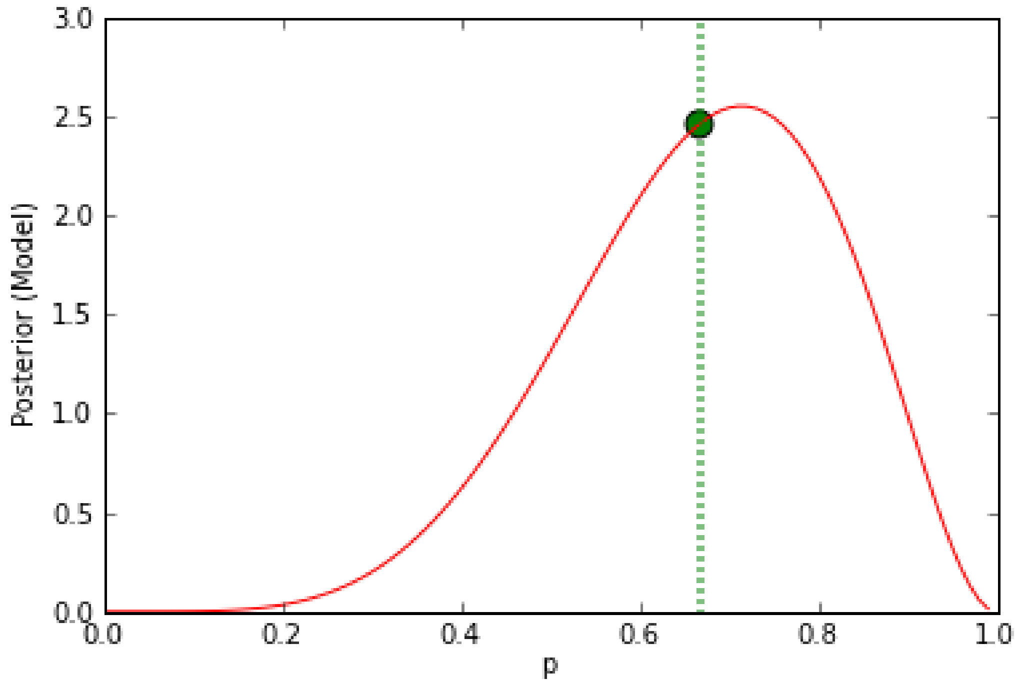
$$E(\text{Model} \mid \text{Data})$$



Bayesian estimation

- Pick the model with minimal *expected risk*

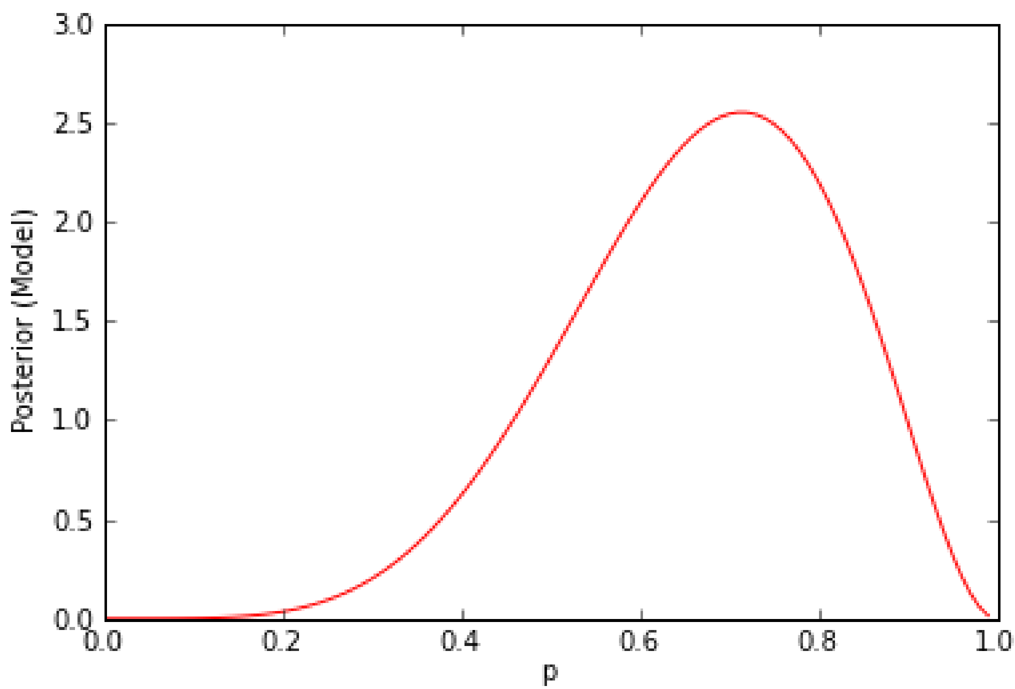
$$E(\text{Model} \mid \text{Data})$$



Bayesian estimation +

- Use the full posterior distribution

$$\Pr(\text{Model} \mid \text{Data})$$



Confidence Intervals

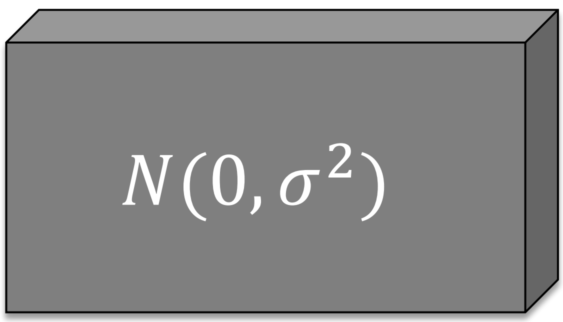
$$\boxed{\hat{p} \pm \frac{1}{\sqrt{N}}}$$

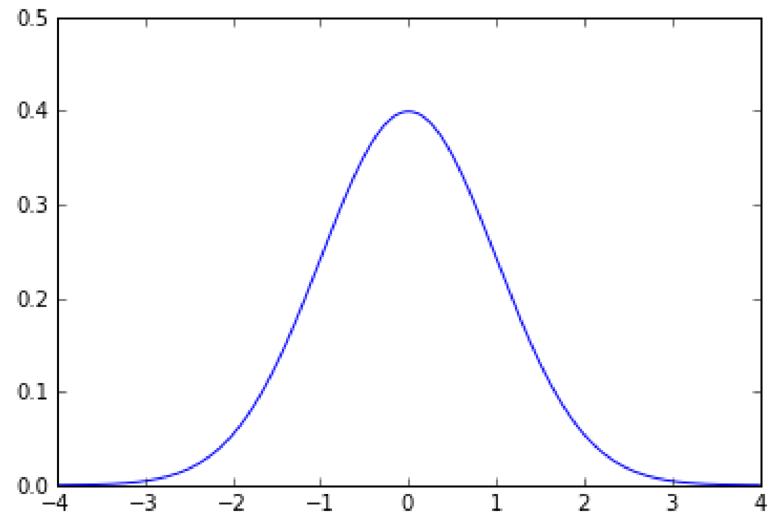
Quiz

- Three major model inference methods are:

Linear Regression (again)

Normal distribution

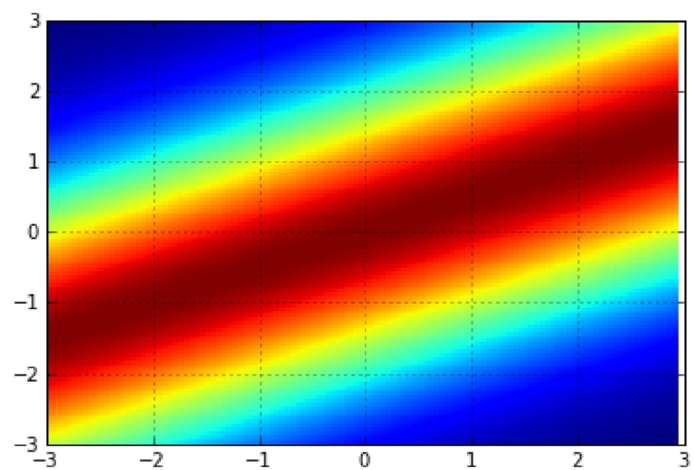




$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right)$$

Linear Regression (again)

$$\boldsymbol{x} \rightarrow \boxed{Y = \boldsymbol{w}^T\boldsymbol{x} + N(0, \sigma^2)}$$



$$Y = \boldsymbol{w}^T \boldsymbol{x} + N(0, \sigma^2)$$

$$Y = \boldsymbol{w}^T \boldsymbol{x} + N(0, \sigma^2)$$

$$e = Y - \boldsymbol{w}^T \boldsymbol{x} \sim N(0, \sigma^2)$$

$$Y = \boldsymbol{w}^T \boldsymbol{x} + N(0, \sigma^2)$$

$$e = Y - \boldsymbol{w}^T \boldsymbol{x} \sim N(0, \sigma^2)$$

$$\Pr((x, y) | \boldsymbol{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{e^2}{\sigma^2}\right)$$

$$\Pr((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \boldsymbol{w}, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$

$$\Pr((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) | \boldsymbol{w}, \sigma^2)$$
$$\propto \prod_i \exp\left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$

$$\log \Pr((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \boldsymbol{w}, \sigma^2)$$
$$\propto \log \prod_i \exp\left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$

$$\log \Pr((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \boldsymbol{w}, \sigma^2)$$
$$\propto \log \prod_i \exp\left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$
$$= \sum_i \left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$

$$\log \Pr((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | \boldsymbol{w}, \sigma^2)$$
$$\propto \log \prod_i \exp\left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$
$$= \sum_i \left(-\frac{1}{2}\frac{e_i^2}{\sigma^2}\right)$$
$$= -\frac{1}{2\sigma^2}\sum_i e_i^2$$

MLE and OLS

$$\operatorname{argmax}_{\mathbf{w}} \mathrm{LogLikelihood}(\mathrm{Data}, \boldsymbol{w}) = \operatorname{argmin}_{\mathbf{w}} \sum_i e_i^2$$

Linear Regression + MAP

$$\Pr(\boldsymbol{w}|\mathrm{Data}) \propto \Pr(\mathrm{Data}|\boldsymbol{w}) \cdot \Pr(\boldsymbol{w})$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\text{Data}) \propto \log \Pr(\text{Data}|\boldsymbol{w}) + \log \Pr(\boldsymbol{w})$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\text{Data}) \propto \boxed{\log \Pr(\text{Data}|\boldsymbol{w})} + \log \Pr(\boldsymbol{w})$$

$$-\sum_i e_i^2$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\text{Data}) \propto \boxed{\log \Pr(\text{Data}|\boldsymbol{w})} + \log \Pr(\boldsymbol{w})$$

$$-\sum_i e_i^2$$

Let the prior on w_j be Gaussian:

$$\Pr(w_j) \propto \exp\left(-\frac{w_j^2}{2\alpha^2}\right)$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\text{Data}) \propto \boxed{\log \Pr(\text{Data}|\boldsymbol{w})} + \log \Pr(\boldsymbol{w})$$

$$-\sum_i e_i^2$$

Let the prior on w_j be Gaussian:

$$\Pr(\boldsymbol{w}) \propto \prod_j \exp\left(-\frac{w_j^2}{2\alpha^2}\right)$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\mathrm{Data}) \propto \boxed{\log \Pr(\mathrm{Data}|\boldsymbol{w})} + \log \Pr(\boldsymbol{w})$$

$$-\sum_i e_i^2$$

Let the prior on w_j be Gaussian:

$$\log \Pr(\boldsymbol{w}) \propto \sum_j -\frac{w_j^2}{2\alpha^2}$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\mathrm{Data}) \propto \boxed{\log \Pr(\mathrm{Data}|\boldsymbol{w})} + \log \Pr(\boldsymbol{w})$$

$$-\sum_i e_i^2$$

Let the prior on w_j be Gaussian:

$$\log \Pr(\boldsymbol{w}) \propto -\sum_j w_j^2$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\text{Data}) \propto \boxed{\log \Pr(\text{Data}|\boldsymbol{w})} + \boxed{\log \Pr(\boldsymbol{w})}$$

$$-\sum_i e_i^2 \qquad -\sum_j w_j^2$$

Let the prior on w_j be Gaussian:

$$\log \Pr(\boldsymbol{w}) \propto -\sum_j w_j^2$$

Linear Regression + MAP

$$\log \Pr(\boldsymbol{w}|\mathrm{Data}) \propto \boxed{\log \Pr(\mathrm{Data}|\boldsymbol{w})} + \boxed{\log \Pr(\boldsymbol{w})}$$

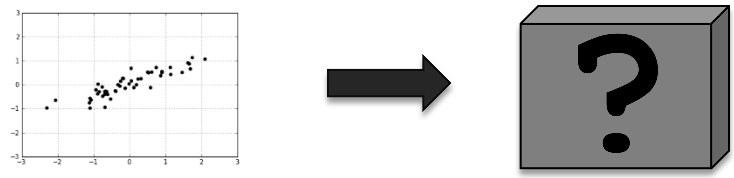
$$-\sum_i e_i^2 \qquad -\sum_j w_j^2$$

Loss ⇔ Error distribution

Penalty ⇔ Model prior

Summary

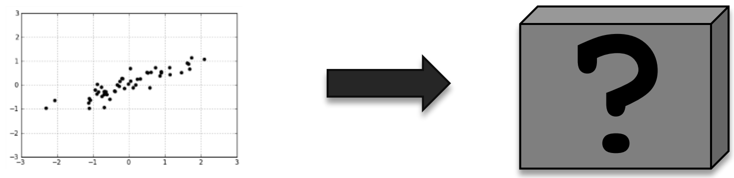
Summary



MLE

MAP

Summary



MLE

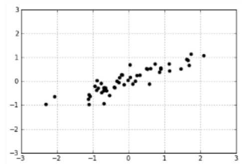
MAP

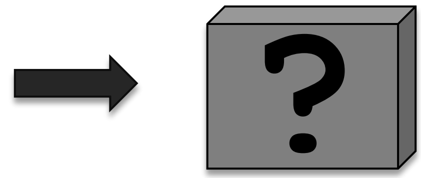
Loss ⇔ Error distribution

Loss ⇔ Error distribution **+**

Penalty⇔ Model prior

Summary





MLE

MAP

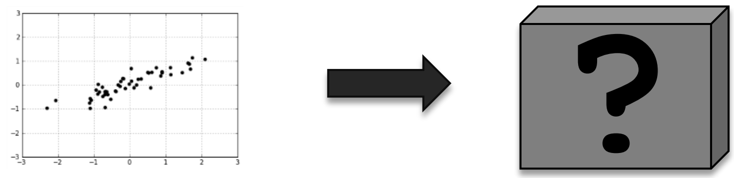
Loss ⇔ Error distribution

Loss ⇔ Error distribution **+** **Penalty⇔ Model prior**

OLS Regression

Ridge Regression

Summary



MLE

MAP

Loss ⇔ Error distribution

Loss ⇔ Error distribution **+**

Penalty⇔ Model prior

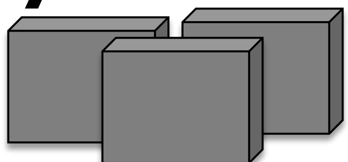
OLS Regression

Ridge Regression

ℓ_2-loss/penalty ⇔ Normal distribution

Summary

- **Probability** for modeling

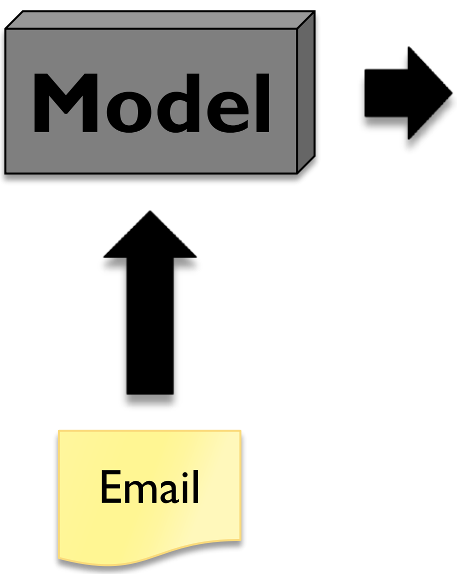


- **Statistics** for estimation

MLE **MAP**

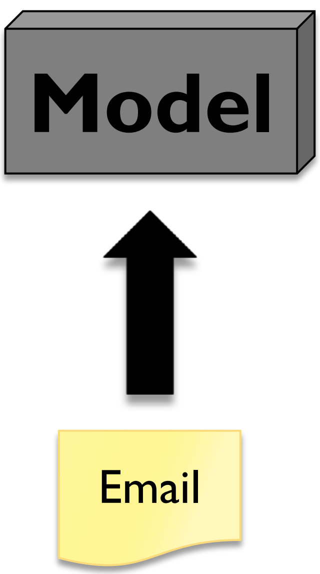
- **Decision theory** for prediction

Decision Theory



X	P(X)
spam	0.8
valid	0.2

Decision Theory

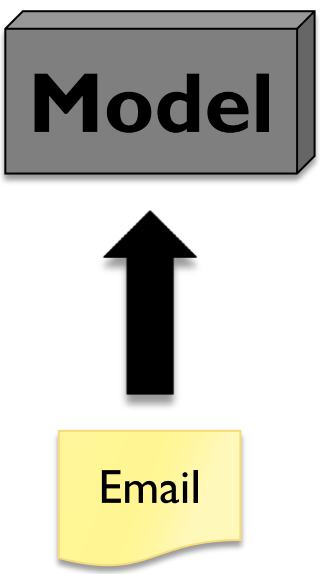


X	P(X)
spam	0.8
valid	0.2

Decision

X	P(X)	SPAM	VALID
spam	0.8		
valid	0.2		

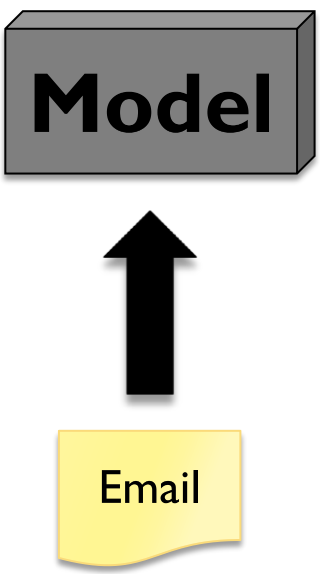
Decision Theory



X	P(X)
spam	0.8
valid	0.2

		Decision	
X	P(X)	SPAM	VALID
spam	0.8	0	1
valid	0.2	5	0

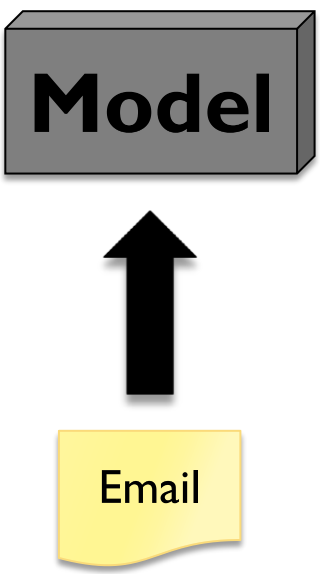
Decision Theory



X	P(X)
spam	0.8
valid	0.2

		Decision	
X	P(X)	SPAM	VALID
spam	0.8	0	1
valid	0.2	5	0
Expected Risk			

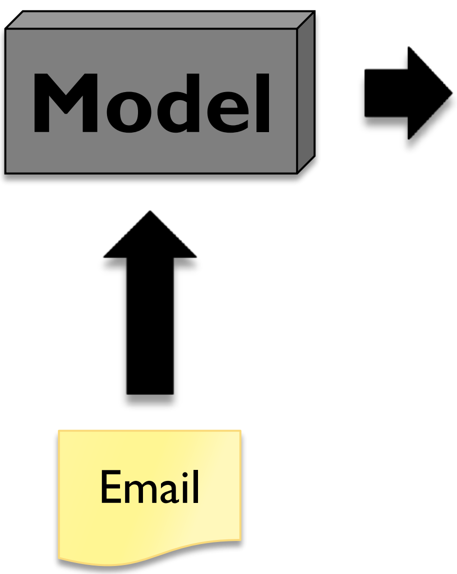
Decision Theory



X	P(X)
spam	0.8
valid	0.2

		Decision	
X	P(X)	SPAM	VALID
spam	0.8	0	1
valid	0.2	5	0
Expected Risk		1	**0.8**

Decision Theory



X	P(X)
spam	0.8
valid	0.2

		Decision	
X	P(X)	SPAM	VALID
spam	0.8	0	1
valid	0.2	5	0
Expected Risk		1	**0.8**

Expected Risk (Supervised Learning)

$$R(\hat{y}|x) = \sum_{y} \Pr(y|x)\ell(\hat{y}, y)$$

Expected Risk (Supervised Learning)

$$R(\hat{y}|x) = \sum_{y} \Pr(y|x)\,\ell(\hat{y}, y)$$

SPAM, VALID

Email

spam, valid

Model

0	1
5	0

Expected Risk (Supervised Learning)

$$R(\hat{y}|x) = \sum_{y} \Pr(y|x)\,\ell(\hat{y}, y)$$

SPAM, VALID

Email

spam, valid

Model

0	**1**
5	**0**

$$R(\hat{y}|x) = \int_{y} \ell(\hat{y}, y)\,dF(y|x)$$

Bayesian Classifier

- Optimal classifier:

For a given x and a conditional probabilistic model

$$\Pr(y|x)$$

predict $\hat{y}$, that has the **smallest expected risk**.

Bayesian Classifier

- Optimal classifier:

For a given x and a conditional probabilistic model

$$\Pr(y|x)$$

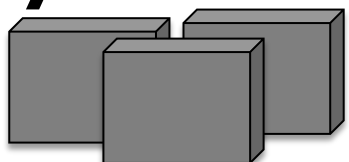
predict $\hat{y}$, that has the **smallest expected risk**.

For *symmetric risk* ℓ, this corresponds to picking the option with the highest probability.

0	1
1	0

Summary

- **Probability** for modeling

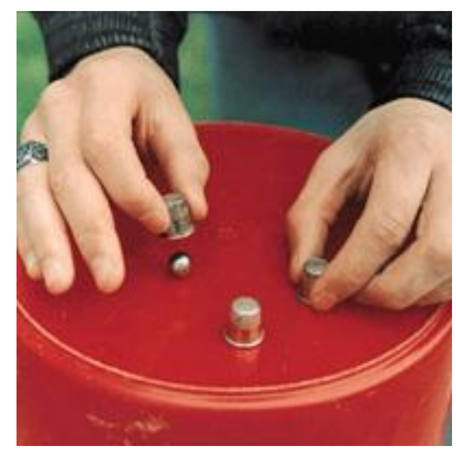


- **Statistics** for estimation

MLE **MAP**

- **Decision theory** for prediction

Bayesian Classifier



The Tennis Dataset

Day	Outlook	Temp	Humidity	Wind	*PlayTennis*
D1	*Sunny*	*Hot*	*High*	*Weak*	*No*
D2	*Sunny*	*Hot*	*High*	*Strong*	*No*
D3	*Overcast*	*Hot*	*High*	*Weak*	*Yes*
D4	*Rain*	*Mild*	*High*	*Weak*	*Yes*
D5	*Rain*	*Cool*	*Normal*	*Weak*	*Yes*
D6	*Rain*	*Cool*	*Normal*	*Strong*	*No*
D7	*Overcast*	*Cool*	*Normal*	*Strong*	*Yes*
D8	*Sunny*	*Mild*	*High*	*Weak*	*No*
D9	*Sunny*	*Cool*	*Normal*	*Weak*	*Yes*
D10	*Rain*	*Mild*	*Normal*	*Weak*	*Yes*
D11	*Sunny*	*Mild*	*Normal*	*Strong*	*Yes*
D12	*Overcast*	*Mild*	*High*	*Strong*	*Yes*
D13	*Overcast*	*Hot*	*Normal*	*Weak*	*Yes*
D14	*Rain*	*Mild*	*High*	*Strong*	*No*

Shall we play tennis today?

PlayTennis
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Shall we play tennis today?

PlayTennis
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Estimate a probabilistic model and predict:

Pr(Yes) = 9/14 = 0.64

Pr(No) = 5/14 = 0.36

➔ Yes

It's windy today. Tennis, anyone?

Wind	PlayTennis
Weak	*No*
Strong	*No*
Weak	*Yes*
Weak	*Yes*
Weak	*Yes*
Strong	*No*
Strong	*Yes*
Weak	*No*
Weak	*Yes*
Weak	*Yes*
Strong	*Yes*
Strong	*Yes*
Weak	*Yes*
Strong	*No*

It's windy today. Tennis, anyone?

Wind	PlayTennis
Weak	*No*
Strong	*No*
Weak	*Yes*
Weak	*Yes*
Weak	*Yes*
Strong	*No*
Strong	*Yes*
Weak	*No*
Weak	*Yes*
Weak	*Yes*
Strong	*Yes*
Strong	*Yes*
Weak	*Yes*
Strong	*No*

Pr(Weak) = 8/14

Pr(Strong) = 6/14

Pr(Yes | Weak) = 6/8

Pr(No | Weak) = 2/8

Pr(Yes | Strong) = 3/6

Pr(No | Strong) = 3/6

More attributes

Humidity	Wind	PlayTennis
High	Weak	No
High	Strong	No
High	Weak	Yes
High	Weak	Yes
Normal	Weak	Yes
Normal	Strong	No
Normal	Strong	Yes
High	Weak	No
Normal	Weak	Yes
Normal	Weak	Yes
Normal	Strong	Yes
High	Strong	Yes
Normal	Weak	Yes
High	Strong	No

Pr(High, Weak) = 4/14

Pr(Yes | High, Weak) = 2/4

Pr(No | High, Weak) = 2/4

Pr(High,Strong) = 3/14

Pr(Yes | High,Strong) = 1/3

Pr(No | High,Strong) = 2/3

…

The Bayesian Classifier

In general:

1. **Estimate from data:**

$$\Pr(\text{Class} \mid x_1, x_2, x_3, \dots)$$

2. **For a given instance** $(x_1, x_2, x_3, \dots)$ **predict class whose conditional probability is greater:**

$$\Pr(\mathrm{C}_1 \mid x_1, x_2, x_3, \dots) > \Pr(\mathrm{C}_2 \mid x_1, x_2, x_3, \dots)$$

$\Rightarrow$ predict C_1

Problem

- **We need exponential amount of data**

Humidity	Wind	PlayTennis
High	*Weak*	*No*
High	*Strong*	*No*
High	*Weak*	*Yes*
High	*Weak*	*Yes*
Normal	*Weak*	*Yes*
Normal	*Strong*	*No*
Normal	*Strong*	*Yes*
High	*Weak*	*No*
Normal	*Weak*	*Yes*
Normal	*Weak*	*Yes*
Normal	*Strong*	*Yes*
High	*Strong*	*Yes*
Normal	*Weak*	*Yes*
High	*Strong*	*No*

Pr(High, Weak) = 4/14

Pr(Yes | High, Weak) = 2/4

Pr(No | High, Weak) = 2/4

Pr(High,Strong) = 3/14

Pr(Yes | High,Strong) = 1/3

Pr(No | High,Strong) = 2/3

…

Naïve Bayes Classifier

To scale beyond 2-3 attributes, use a hack:

***Assume* that attributes are independent within each class:**

$$\Pr(x_1, x_2, x_3 \mid \text{Class}) \\ = \Pr(x_1|\text{Class})\Pr(x_2|\text{Class})\Pr(x_3|\text{Class}) \dots$$

Naïve Bayes Classifier

1. $\Pr(C_1|\boldsymbol{x}) > \Pr(C_2|\boldsymbol{x})$
   ➔ predict C_1

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

Naïve Bayes Classifier

1. $\Pr(C_1|\boldsymbol{x}) > \Pr(C_2|\boldsymbol{x})$
   $\rightarrow$ predict C_1
2. $\dfrac{\Pr(C_1)\Pr(\boldsymbol{x}|C_1)}{\Pr(\boldsymbol{x})} > \dfrac{\Pr(C_2)\Pr(\boldsymbol{x}|C_2)}{\Pr(\boldsymbol{x})}$
   $\rightarrow$ predict C_1

Naïve Bayes Classifier

1. $\Pr(\mathrm{C}_1|\boldsymbol{x}) > \Pr(\mathrm{C}_2|\boldsymbol{x})$

   ➔ predict C_1

2. $$\frac{\Pr(\mathrm{C}_1)\Pr(\boldsymbol{x}|\mathrm{C}_1)}{\Pr(\boldsymbol{x})} > \frac{\Pr(\mathrm{C}_2)\Pr(\boldsymbol{x}|\mathrm{C}_2)}{\Pr(\boldsymbol{x})}$$

   ➔ predict C_1

3. $\Pr(\mathrm{C}_1)\Pr(\boldsymbol{x}|\mathrm{C}_1) > \Pr(\mathrm{C}_2)\Pr(\boldsymbol{x}|\mathrm{C}_2)$

   ➔ predict C_1

Naïve Bayes Classifier

1. $\Pr(C_1|\boldsymbol{x}) > \Pr(C_2|\boldsymbol{x})$
   ➔ predict C_1
2. $\dfrac{\Pr(C_1)\Pr(\boldsymbol{x}|C_1)}{\Pr(\boldsymbol{x})} > \dfrac{\Pr(C_2)\Pr(\boldsymbol{x}|C_2)}{\Pr(\boldsymbol{x})}$
   ➔ predict C_1
3. $\Pr(C_1)\Pr(\boldsymbol{x}|C_1) > \Pr(C_2)\Pr(\boldsymbol{x}|C_2)$
   ➔ predict C_1
4. $\Pr(C_1) \cdot \Pr(x_1|C_1)\Pr(x_2|C_1)\ldots\Pr(x_m|C_1) >$
   $\Pr(C_2) \cdot \Pr(x_1|C_2)\Pr(x_2|C_2)\ldots\Pr(x_m|C_2)$
   ➔ predict C_1

Naïve Bayes Classifier

1. $\Pr(C_1|\boldsymbol{x}) > \Pr(C_2|\boldsymbol{x})$
   $\rightarrow$ predict C_1
2. $\dfrac{\Pr(C_1)\Pr(\boldsymbol{x}|C_1)}{\Pr(\boldsymbol{x})} > \dfrac{\Pr(C_2)\Pr(\boldsymbol{x}|C_2)}{\Pr(\boldsymbol{x})}$
   $\rightarrow$ predict C_1
3. $\Pr(C_1)\Pr(\boldsymbol{x}|C_1) > \Pr(C_2)\Pr(\boldsymbol{x}|C_2)$
   $\rightarrow$ predict C_1
4. $\boxed{\begin{matrix}\Pr(C_1)\\ \Pr(C_2)\end{matrix}} \begin{matrix}\cdot\\ \cdot\end{matrix} \boxed{\begin{matrix}\Pr(x_1|C_1)\Pr(x_2|C_1)\ldots\Pr(x_m|C_1)\\ \Pr(x_1|C_2)\Pr(x_2|C_2)\ldots\Pr(x_m|C_2)\end{matrix}} >$
   $\rightarrow$ predict C_1

Naïve Bayes Classifier

- Works for both discrete and continuous attributes.

- The goods:
  - Easy to implement, efficient
  - Won't overfit, intepretable
  - Works better than you would expect (e.g. spam filtering)
- The bads
  - "Naïve", linear
  - Usually won't work well for too many classes
  - Not a good probability estimator

Naïve Bayes Classifier

```
from sklearn.naive_bayes import
                       BernoulliNB,
                       MultinomialNB,
                       GaussianNB
```

Quiz

- MLE:

$$\mathrm{argmax}_{\mathrm{Model}}\underline{\hspace{6cm}}$$

- MAP:

$$\mathrm{argmax}_{\mathrm{Model}}\underline{\hspace{6cm}}$$

- Gaussian distribution:

$$f(x) = \mathrm{const} \times \exp(\underline{\hspace{2.5cm}})$$

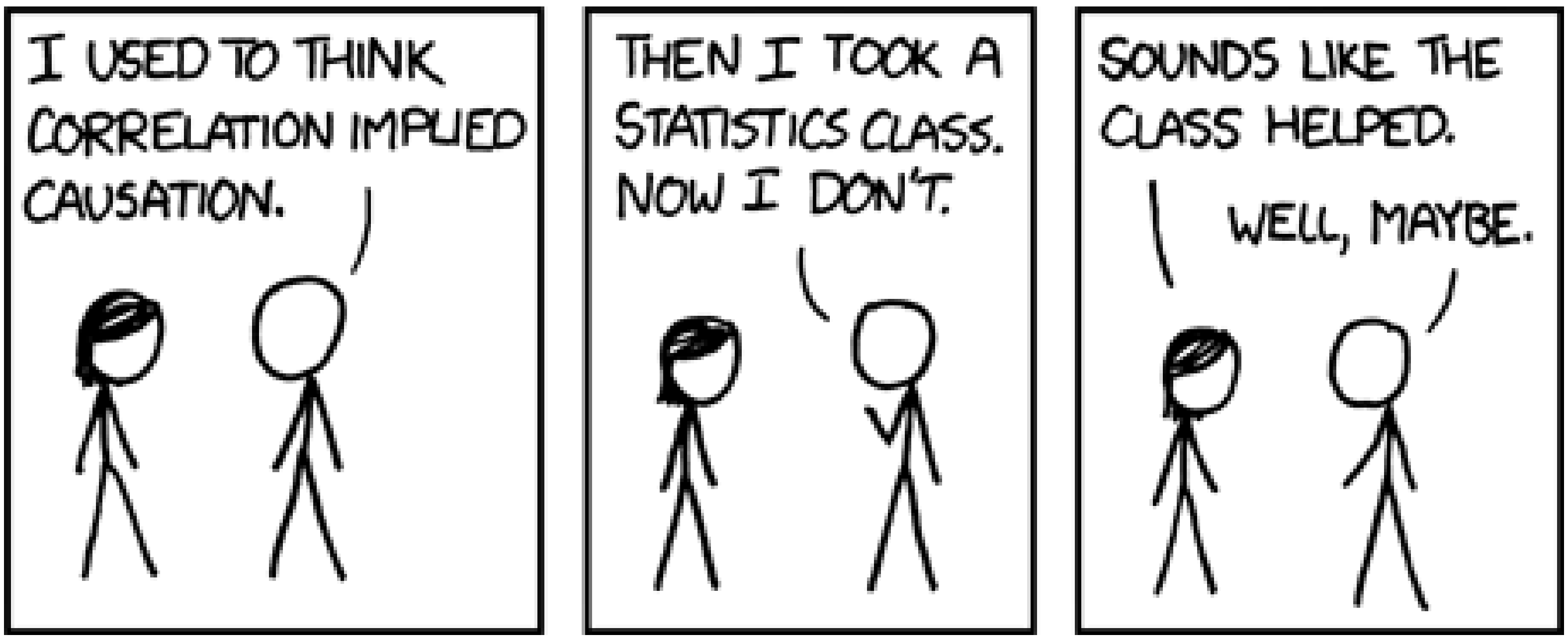
Quiz

- Bayesian classifier has optimal ________________

- Naïve Bayesian classifier assumption:
  - $\Pr(C|x_1, x_2) = \Pr(C|x_1)\Pr(C|x_2)$
  - $\Pr(x_1, x_2|C) = \Pr(x_1|C)\Pr(x_2|C)$
  - $\Pr(C_1, C_2|x) = \Pr(C_1|x)\Pr(C_2|x)$
  - $\Pr(x|C_1, C_2) = \Pr(x|C_1)\Pr(x|C_2)$

- All machine learning methods we have considered so far rely on MLE or MAP
  - Yes
  - No

Questions?



http://xkcd.com/552/