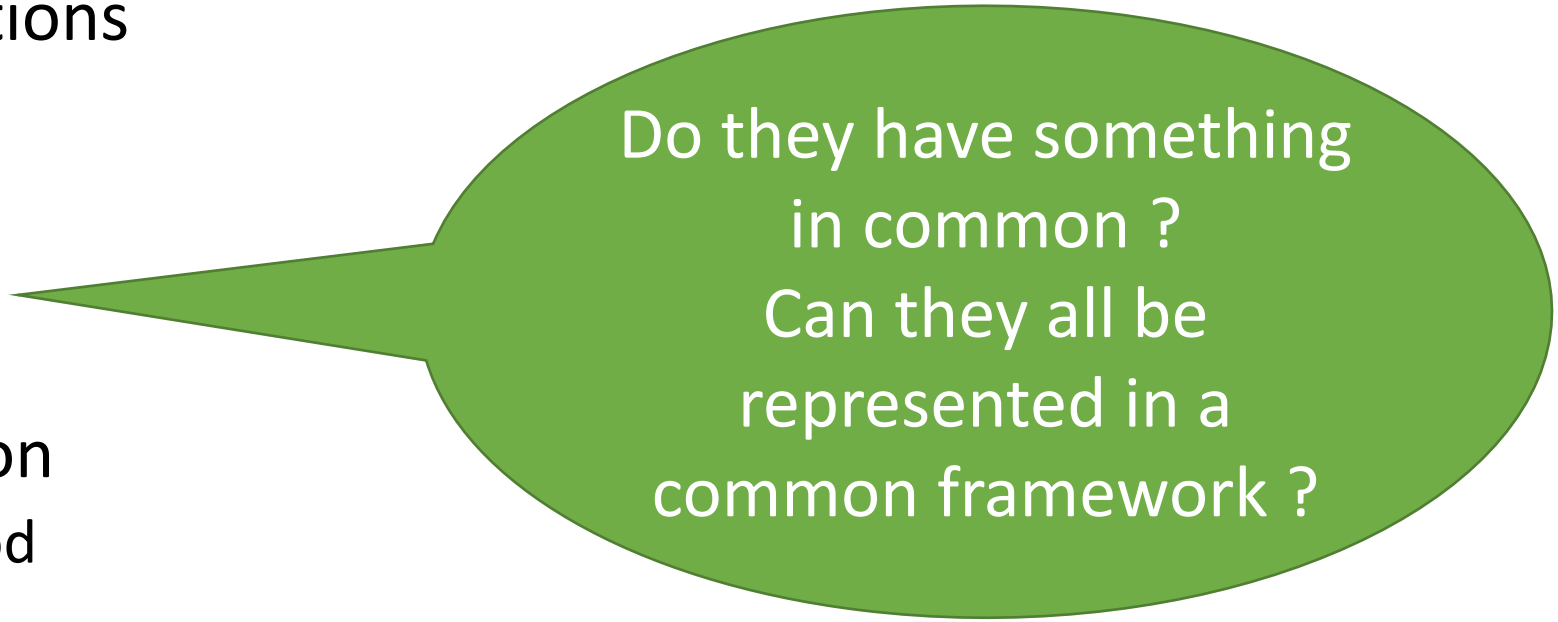
Parameter Estimation and Inference

Recap

- Probability Distributions
  - Bernoulli
  - Multinomial
  - Poisson
  - Gaussian
- Parameter Estimation
  - Maximum Likelihood

Today

- Probability Distributions
  - Bernoulli
  - Multinomial
  - Poisson
  - Gaussian
- Parameter Estimation
  - Maximum Likelihood



Today

- Probability Distributions
  - Bernoulli
  - Multinomial
  - Poisson
  - Gaussian
- Parameter Estimation
  - Maximum Likelihood

Do they have something in common ?

What are the properties of MLE?
Does it have some disadvantage ?
Does there exist other estimates ?

Exponential Family

Can Bernoulli distribution be represented in exponential form ?

Bernoulli distribution

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}.$$

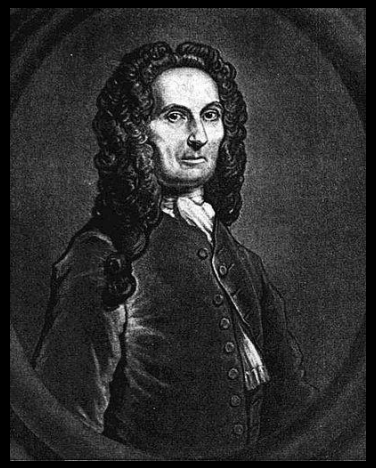
$$\begin{aligned} p(x|\mu) &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}. \end{aligned}$$

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

Exponential form

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \qquad \sigma(\eta) = \frac{1}{1+\exp(-\eta)}$$

Exponential Family



Exponential Family

- Bernoulli distribution : $p(x|\eta) = \sigma(-\eta)\exp(\eta x)$

- Exponential family of distributions over **x**, given parameters $\boldsymbol{\eta}$,

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

$\boldsymbol{\eta}$ are called the *natural parameters* of the distribution, and **u(x)** is some function of **x**.

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

// ensures that the distribution is normalized

Exponential Family

Exponential family of distributions over **x**, given parameters ***η***

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

Bernoulli distribution is exponential Family !

$$p(x|\eta) = \sigma(-\eta) \exp(\eta x)$$

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta). \end{aligned}$$

Gaussians as exponential family

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \end{aligned}$$

Expectation parameters

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \\ \mathbf{u}(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ h(\mathbf{x}) &= (2\pi)^{-1/2} \\ g(\boldsymbol{\eta}) &= (-2\eta_2)^{1/2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right). \end{aligned}$$

Natural parameters

Maximum likelihood estimation in exponential family

- Estimating the parameter vector ***η*** *using maximum likelihood*

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \qquad g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

- Taking derivative wrt **η**

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x}$$

$$+ \quad g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0.$$

$$-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

$$\boxed{-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})].}$$

Maximum likelihood and sufficient statistics

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

- Maximum likelihood estimate of **η** can be estimated based on sum of u(**x**) over data points which is called the *sufficient statistic* of distribution.
- No not need to store the entire data set itself but only the value of the sufficient statistic
- For the Gaussian **u**(*x*) = (*x, x*^2), and so we should keep both the sum of *{xn}* and the sum of *{xn^2}*.

Parameter Estimation : Method of Moments

- Method of moments is a method of estimation of population parameters.
- Moments of a random variable

E[X]	First Moment (Mean)
E[X^2]	Second moment
E[X^n]	nth moment of X

- Parameters of distribution can be expressed in terms of moments (i.e., the expected values of powers of the random variable under consideration. Those expressions are then set equal to the sample moments computed from data.

Parameter Estimation : Method of Moments

- Example 1 [**Bernoulli**] Let X1, . . . , Xn ∼ Bernoulli(p). p represents expected value of random variable. Hence p = E[X]

$$\alpha_1 = \mathbb{E}_p(X) = p$$

Moment as a function of parameters

$$\widehat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

sample moment

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Estimate of p

Parameter Estimation : Method of Moments

- Example 2 [**Gaussian**] Let X1, . . . , Xn ∼ Normal(μ, σ2).

$$\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$$

$$\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2.$$

Moment as a function of parameters

Parameter Estimation : Method of Moments

- Example 2 [**Gaussian**] Let X1, . . . , Xn ∼ Normal(μ, σ2).

$$\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$$

$$\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2.$$

Moment as a function of parameters

$$\widehat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\mu}$$

$$\widehat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2. = \widehat{\sigma}^2 + \widehat{\mu}^2$$

sample moments

Estimate of μ, σ2 by solving the system of equations

Parameter Estimation : Method of Moments

- Example 2 [**Gaussian**] Let X1, . . . , Xn ~ Normal(μ, σ2).

sample moments

$$\widehat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\mu}$$

$$\widehat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2. = \widehat{\sigma}^2 + \widehat{\mu}^2$$

Estimate of μ, σ2 by solving the system of equations

$$\widehat{\mu} = \overline{X}_n \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Parameter Estimation : Method of Moments

The **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned}$$

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) \qquad \widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

defines a system of k equations with k unknowns.

EVALUATING A POINT ESTIMATOR

- let d = d (X) be an estimator of θ.
- Mean square error of the estimator d

$$r(d,\theta) = E[(d(\mathbf{X}) - \theta)^2]$$

- bias of d as an estimator of θ. $b_\theta(d) = E[d(\mathbf{X})] - \theta$
- bθ(d) = 0 for all θ, then we say that d is an unbiased estimator of θ.
- An estimator is unbiased if its expected value always equals the value of the parameter

Bias of an estimator

- Consider estimator of population mean

$$d_2(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \theta$$

$$, d_3(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \lambda_i X_i \qquad \sum_{i=1}^{n} \lambda_i = 1.$$

$$E\left[\sum_{i=1}^{n} \lambda_i X_i\right] = \sum_{i=1}^{n} E[\lambda_i X_i] \quad = \sum_{i=1}^{n} \lambda_i E(X_i) \quad = \theta \sum_{i=1}^{n} \lambda_i$$

- How about estimator of population variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \mathbf{E}[S^2] = \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2.$$

Mean square error, Bias, Variance

- Mean square error of unbiased estimator

$$r(d,\theta) = E[(d(\mathbf{X}) - \theta)^2] \quad = E[(d(\mathbf{X}) - E[d(\mathbf{X})])^2] = \mathrm{Var}(d(\mathbf{X}))$$

- mean square error of any estimator is equal to its variance plus the square of its bias.

$$r(d,\theta) = \mathrm{Var}(d) + b_\theta^2(d)$$

MSE Variance Bias

$$
\begin{aligned}
r(d,\theta) &= E[(d(\mathbf{X}) - \theta)^2] \\
&= E[(d - E[d] + E[d] - \theta)^2] \\
&= E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(E[d] - \theta)(d - E[d])] \\
&= E[(d - E[d])^2] + E[(E[d] - \theta)^2] \\
&\quad + 2E[(E[d] - \theta)(d - E[d])] \\
&= E[(d - E[d])^2] + (E[d] - \theta)^2 + 2(E[d] - \theta)E[d - E[d]] \\
&\qquad\qquad \text{since } E[d] - \theta \text{ is constant} \qquad\qquad E[d - E[d]] = 0 \\
&= E[(d - E[d])^2] + (E[d] - \theta)^2
\end{aligned}
$$

$$
r(d,\theta) = \mathrm{Var}(d) + b_\theta^2(d)
$$

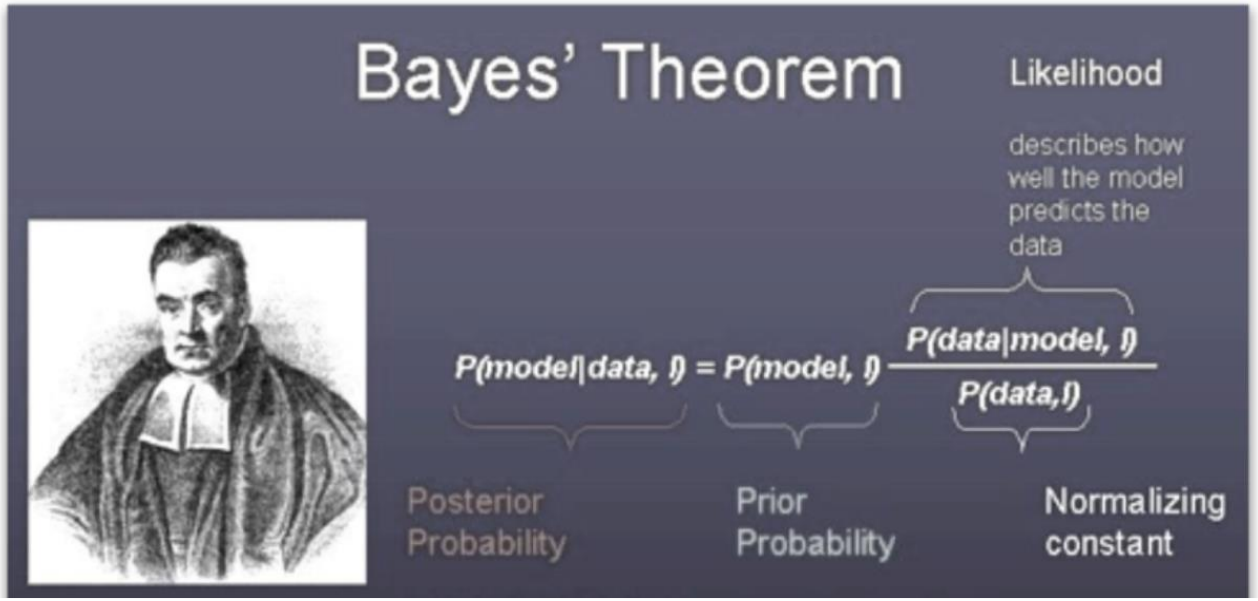
Maximum Likelihood Estimate

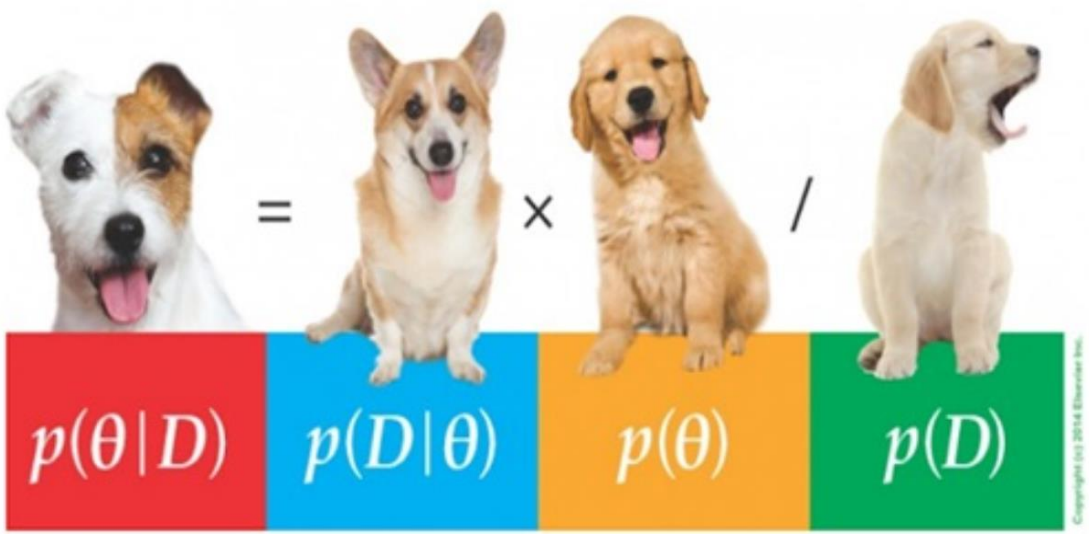
- Rain prediction
- Observed data [1,1,1,1,1,1,1,1,1,1]
- What’s the probability that it will rain in some day in future ?

Bayes estimator

- Assume its summer, and you have collected past 10 days data where it didn't rain a single day. Whats the MLE of θ, the probability that it rains some day ?
- we have some information (prior)about the value of θ, treat θ as a random variable and express prior information as a probability distribution.
- θ is equally likely to be near any value in the interval (0, 1), θ is chosen from a uniform distribution on (0, 1).
- f (x|θ) represents the likelihood that observed data value is equal to x when θ is the value of the parameter.
- How will you combine the prior information about θ with the observed information in X ?

Bayes Theorem





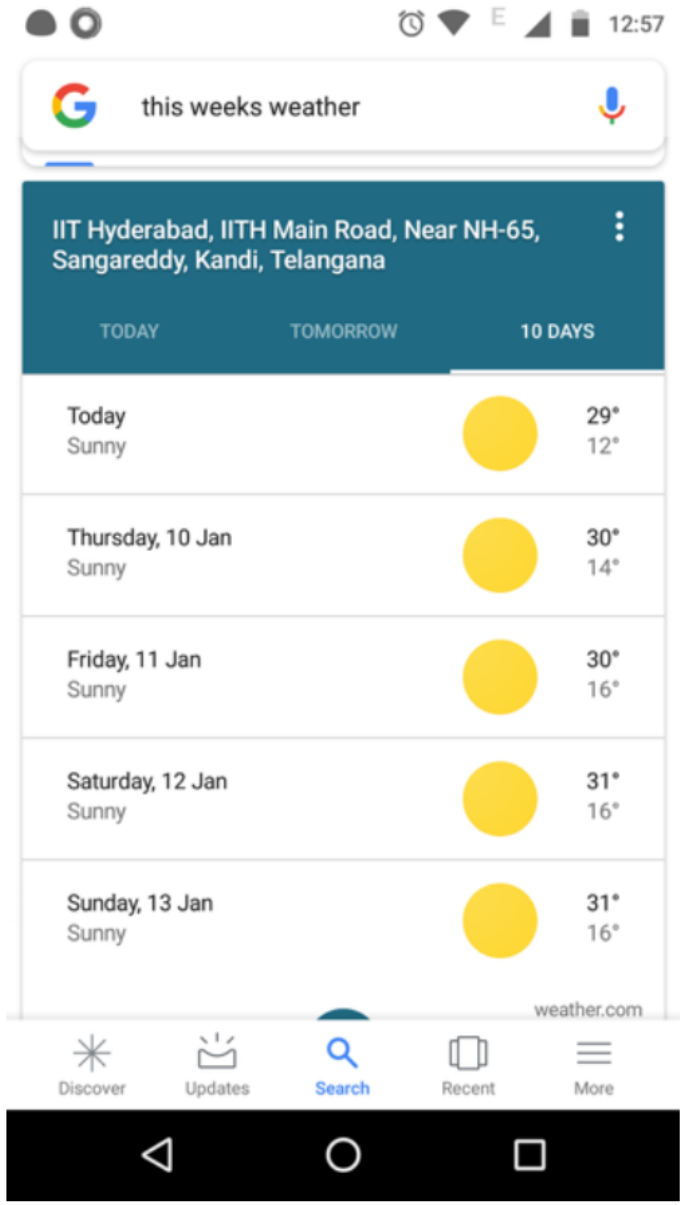
Rain prediction : ML estimation

X = [1,0,0, 0, 1, 0,….]

Model ?
Likelihood : p(X l model)
Learn model parameters : maximum likelihood (ML) estimation

$$\widehat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\ \mathcal{L}$$

$$\mathcal{L} = \sum_{\mathbf{x}_i \in \mathcal{X}} log\ prob(\mathbf{x}_i|\Theta)$$



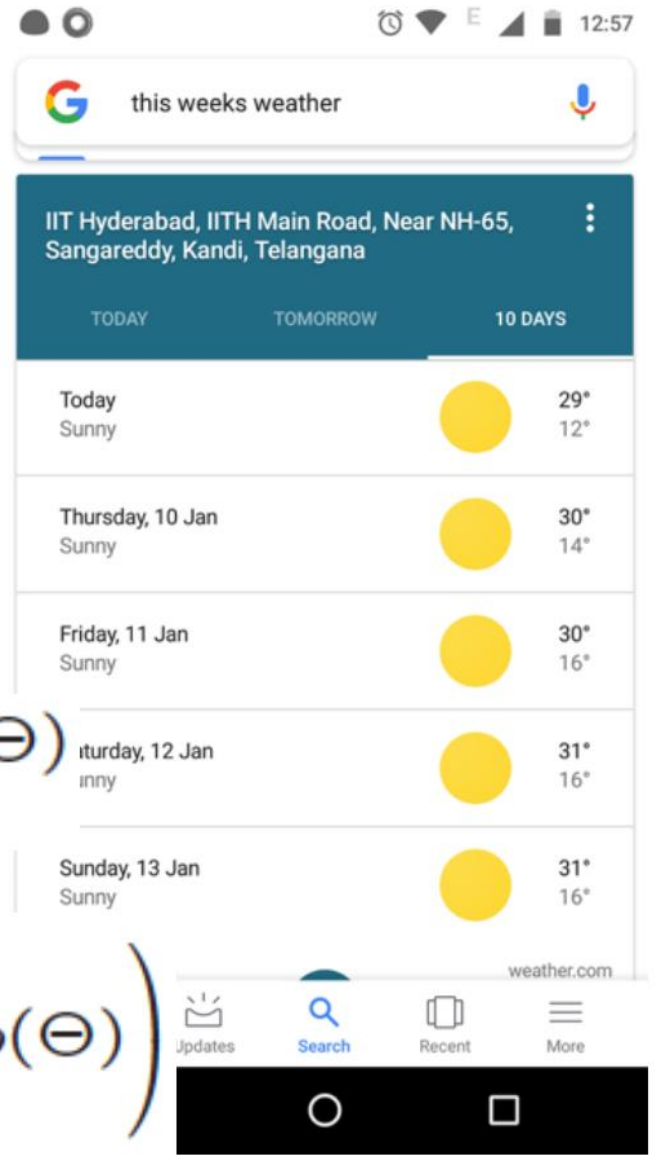
Rain prediction : MAP estimation

Maximum A-Posteriori (MAP) estimation

$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})}$$

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta) \cdot prob(\Theta)$$

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left(\sum_{\mathbf{x}_i \in \mathcal{X}} log\ prob(\mathbf{x}_i|\Theta) + log\ prob(\Theta) \right)$$



Seek that value for θ which maximizes the posterior prob(θ|X).

What Does the MAP Estimate Get Us That the ML Estimate Does NOT

- MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in θ
- MAP estimation "pulls" the estimate toward the prior. The more focused our prior belief, the larger the pull toward the prior.
- "smoothing" role (Laplace smoothing) for parameter estimation.

MAP Estimation Example [US Election]

- Consider a survey where voters were asked if they will vote for NDA or UPA in the next election. Let p be the probability that an individual will vote NDA.
- xi is either NDA or UPA,
- nd is the number of individuals who a planning to vote NDA

$$\mathcal{X} = \left\{ x_i = \begin{vmatrix} Democratic \\ \\ Republican \end{vmatrix}, \quad i = 1...N \right\}$$

$$\widehat{p}_{ML} = \frac{n_d}{N}$$

If N = 20 and if 12 out of 20 said that they were going to vote NDA, we get the following the ML estimate for p: ˆpML = 0.6.

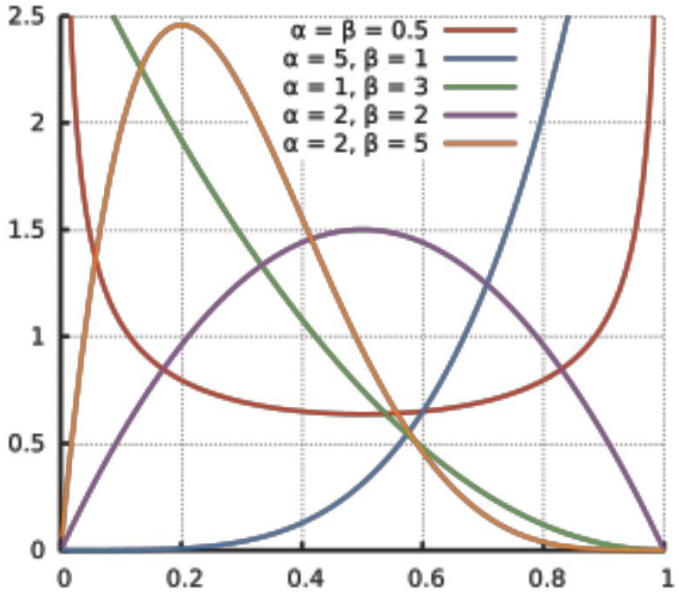
MAP Estimate : Prior Belief on p

- The prior for p must be zero outside the [0, 1] interval.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the [0, 1] interval.
- beta distribution that is parameterized by two "shape" constants α and β does the job nicely for expressing our prior beliefs concerning p:

$$prob(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$B = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\text{mode} \quad \frac{\alpha-1}{\alpha+\beta-2}$$



Example [Indian Election]

- The state of Kerala has traditionally voted UPA. However, on account of the prevailing conditions, the voters are more likely to vote NDA in the election in question.
- a prior distribution for p that has a peak at 0.5. Setting α = β=5 gives us a distribution for p that has a peak in the middle of the [0, 1] interval.

$$\hat{p}_{MAP} = \underset{p}{\text{argmax}} \left(\sum_{x \in \mathcal{X}} log\ prob(x|p) \ + \ log\ prob(p) \right)$$

MAP Estimation [Indian Election]

$$\hat{p}_{MAP} = \underset{p}{\text{argmax}} \Big(n_d \cdot log\ p + (N - n_d) \cdot log\ (1 - p) + log\ prob(p) \Big)$$

$$\frac{n_d}{p} - \frac{(N - n_d)}{(1 - p)} + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1 - p} = 0$$

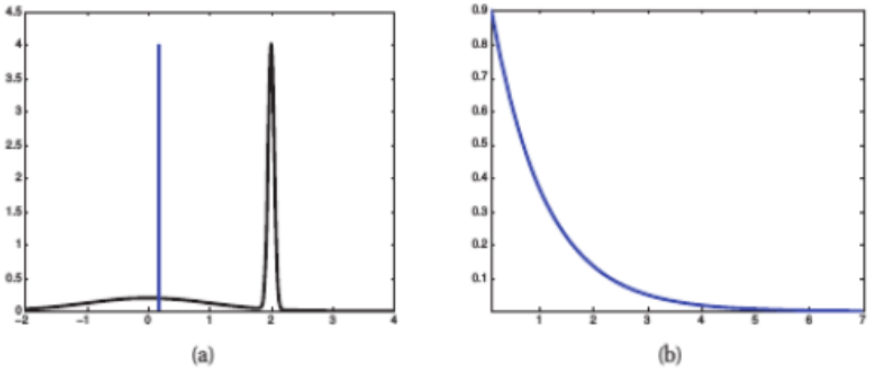
$$\hat{p}_{MAP} = \frac{n_d + \alpha - 1}{N + \alpha + \beta - 2}$$

$$= \frac{n_d + 4}{N + 8}$$

With N = 20 and with 12 of the 20 saying they would vote Democratic, the MAP estimate for p is 0.571 with α and β both set to 5.

Rain prediction : Bayesian estimation

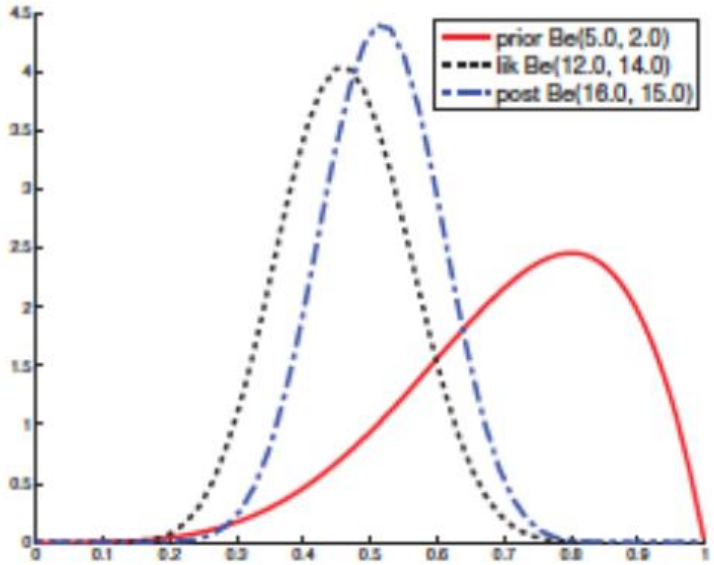
Both ML and MAP return only single and specific values for the parameter Θ.



$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})}$$

$$prob(\mathcal{X}) = \int_{\Theta} prob(\mathcal{X}|\Theta) \cdot prob(\Theta)\ d\Theta$$

$$prob(\tilde{\mathbf{x}}|\mathcal{X}) = \int_{\Theta} prob(\tilde{\mathbf{x}}|\Theta) \cdot prob(\Theta|\mathcal{X})\ d\Theta$$



Posterior is a “compromise” between the prior and likelihood.

posterior mean is convex combination of the prior mean and the MLE

: $\lambda m_1 + (1-\lambda)\hat{\theta}_{MLE}$

Bayesian Estimation and Prediction

X : BINARY R.V.

$X \sim$ BERNOULLI (q

PRIOR

$p(q) \sim \text{UNIF}(0,1)$

Bayesian Estimation and Prediction

X : BINARY R.V $P(x|q) = q^x (1-q)^{1-x}$

$X \sim$ BERNOULLI (q

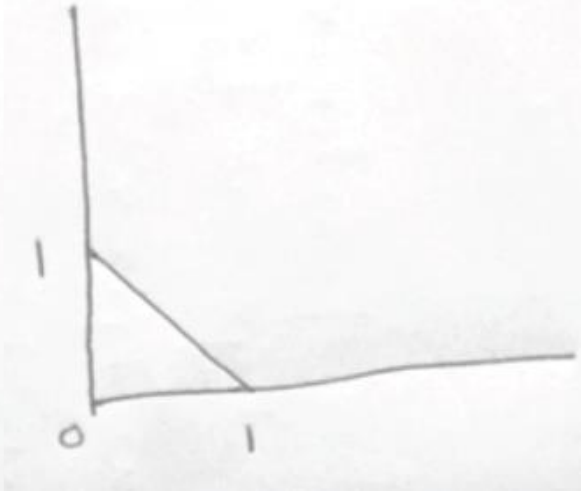
LIKELIHOOD

$P(x=0|q)$

$= 1-q$

PRIOR

$P(q) \sim \text{UNIF}(0,1)$



Bayesian Estimation and Prediction

X : BINARY R.V $P(x|q) = q^x (1-q)^{1-x}$

$X \sim$ BERNOULLI (q

PRIOR

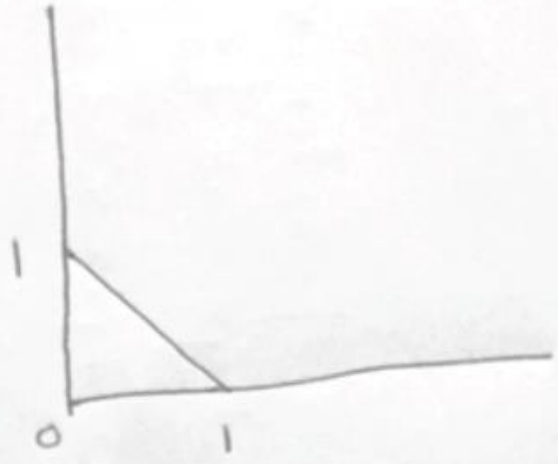
$P(q) \sim \text{UNIF}(0,1)$

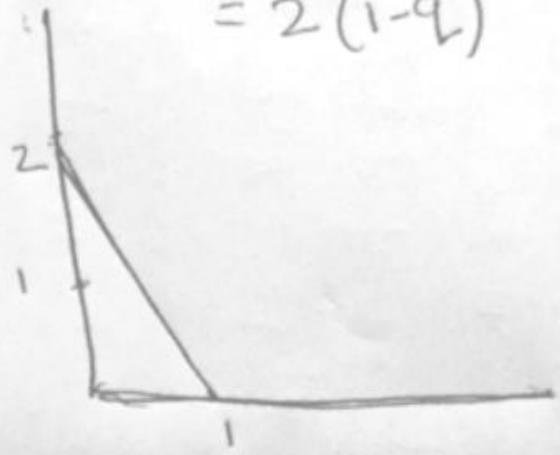
LIKELIHOOD

$P(x=0|q) = 1-q$

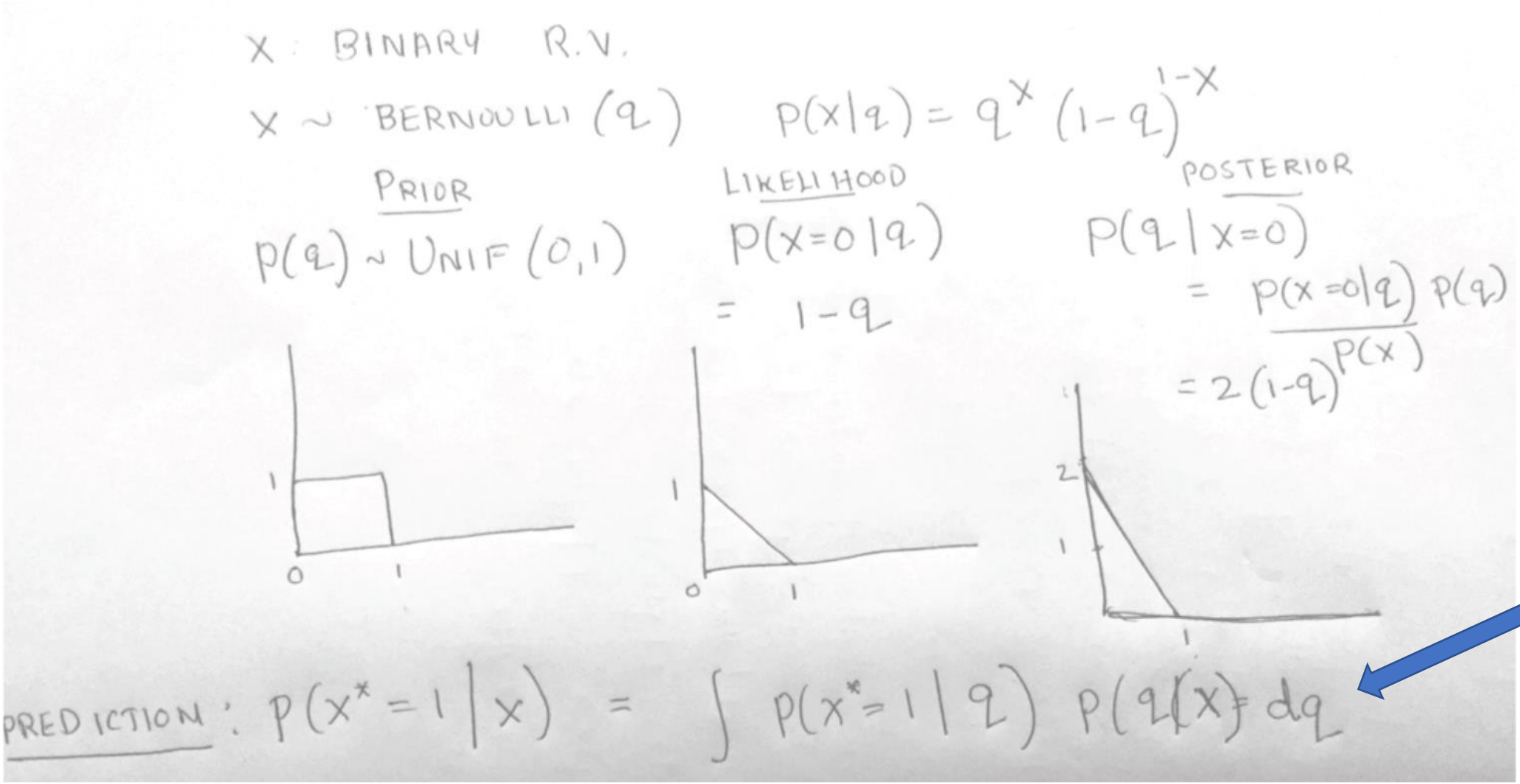
POSTERIOR

$P(q|x=0) = \frac{P(x=0|q)\,P(q)}{P(x)} = 2(1-q)$





Bayesian Estimation and Prediction



Bayes Estimator

- Bayes Estimator : best estimate of θ, given the data values mean of the posterior distribution f (θ|x1, . . . , xn).

$$E[\theta|X_1 = x_1, \ldots, X_n = x_n] = \int \theta f(\theta|x_1, \ldots, x_n)\, d\theta$$

- Mean of posterior over θ, the best estimate of the value of that random variable, in the sense of minimizing the expected squared error

Bayes Estimator

- Suppose that the value of a random variable X is to be predicted.
- If we predict that X will equal c, then the square of the “error” involved will be (X – c)^2.

$$
\begin{aligned}
E[(X-c)^2] &= E[(X-\mu+\mu-c)^2] \\
&= E[(X-\mu)^2 + 2(\mu-c)(X-\mu) + (\mu-c)^2] \\
&= E[(X-\mu)^2] + 2(\mu-c)E[X-\mu] + (\mu-c)^2 \\
&= E[(X-\mu)^2] + (\mu-c^2) \quad \text{since} \quad E[X-\mu] = E[X] - \mu = 0 \\
&\geq E[(X-\mu)^2]
\end{aligned}
$$

average squared error is minimized when we predict that X will equal its mean μ.

Bayesian Machine Learning

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	likelihood of θ
$P(\theta)$	prior probability of θ
$P(\theta\|\mathcal{D})$	posterior of θ given $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$