Lecture 2. Statistical Schools of Thought

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio

Frequentist statistics

Independent and identically distributed

- Abstract problem
  - Given: $X_1, X_2, \ldots, X_n$ drawn i.i.d. from some distribution
  - Want to: identify unknown distribution, or a property of it
- Parametric approach ("**parameter estimation**")
  - Class of models $\{p_\theta(x): \theta \in \Theta\}$ indexed by parameters Θ (could be a real number, or vector, or ….)
  - Point estimate $\hat{\theta}\,(X_1, \ldots, X_n)$ a function (or statistic) of data

Hat means estimate or estimator

- Examples
  - Given *n* coin flips, determine probability of landing heads
  - Learning a classifier

Estimator Bias

Frequentists seek good behaviour, in ideal conditions

- Bias: $B_\theta(\hat{\theta}) = E_\theta\left[\hat{\theta}(X_1, \dots, X_n)\right] - \theta$

Subscript θ means data *really* comes from p_θ

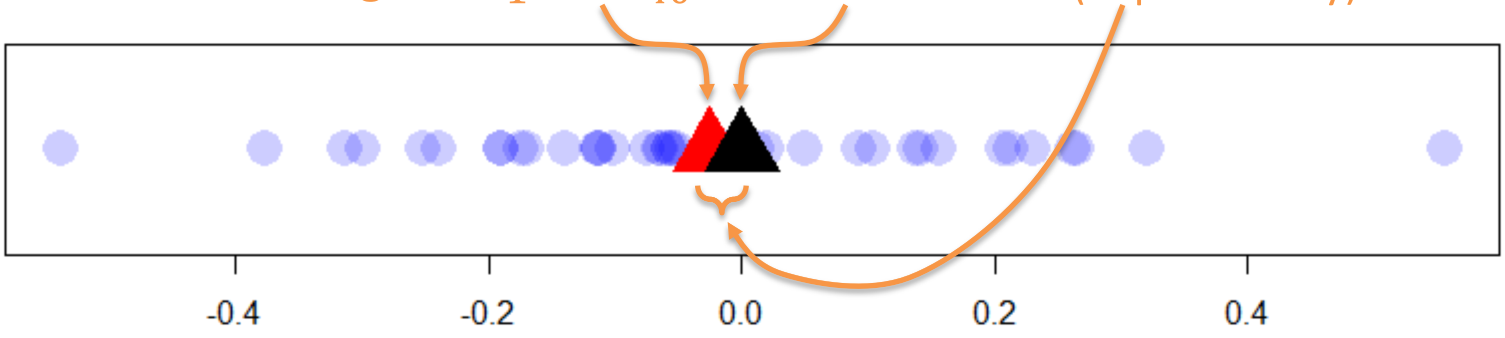
Example: for *i*=1…40

- $X_{i,1}, \dots, X_{i,20} \sim p_\theta = Normal(\theta = 0, \sigma^2 = 1)$
- $\hat{\theta}_i = \frac{1}{20}\sum_{j=1}^{20} X_{i,j}$ the sample mean, plot as ●

Average of $\hat{\theta}_1, \dots, \hat{\theta}_{40}$

True $\theta = 0$

Gap converges to bias (in probability)!



Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}])^2\right]$

$\hat{\theta}$ still function of data

Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[\left(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}]\right)^2\right]$

$\hat{\theta}$ still function of data

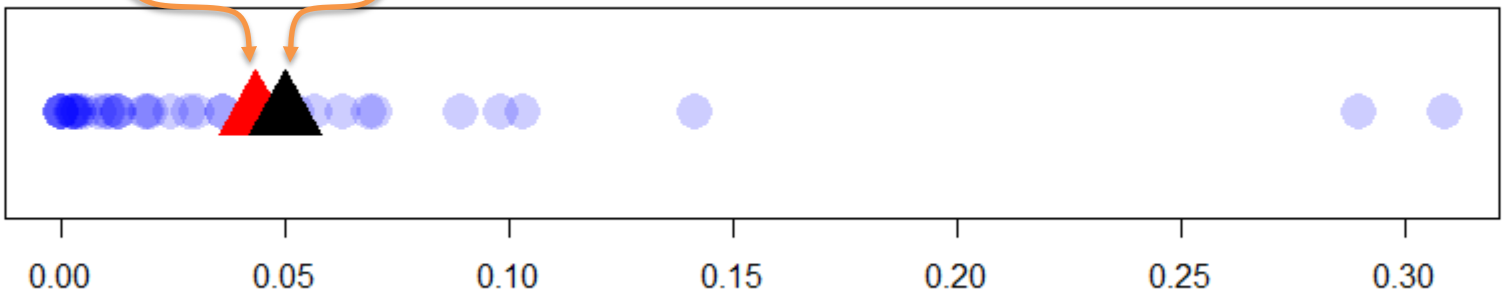
Example cont.

- Plot each $(\hat{\theta}_i - \mathrm{E}_\theta[\hat{\theta}_i])^2 = \hat{\theta}_i^{\,2}$ as ●

Average of $\hat{\theta}_1^{\,2}, \ldots, \hat{\theta}_{40}^{\,2}$

True $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \frac{\sigma^2}{20} = 0.05$

Once again, average converges to true (in probability)!



Asymptotically Well Behaved

For our example estimator (sample mean), we could calculate its exact bias (zero) and variance (σ^2). Usually can't guarantee low bias/variance exactly ☹

Asymptotic properties often hold! ☺

- Consistency: $\hat{\theta}(X_1, \ldots, X_n) \to \theta$ in probability
- Asymptotic efficiency: $\mathrm{Var}_\theta\left(\hat{\theta}(X_1, \ldots, X_n)\right)$ converges to the smallest possible variance of any estimator of θ

Bias closer and closer to zero

Variance closer & closer to optimal

Maximum-Likelihood Estimation

- A general principle for designing estimators
- Involves optimisation
- $\hat{\theta}(x_1, \dots, x_n) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(x_i)$
- *"The best estimate is one under which observed data is most likely"*

Fischer

Example I: Bernoulli

- Know data comes from Bernoulli distribution with unknown parameter (e.g., biased coin); find mean
- MLE for mean
  - $p_\theta(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases} = \theta^x (1-\theta)^{1-x}$
    (note: $p_\theta(x)$ = 0 for all other x)
  - Maximising likelihood yields $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Derivation: https://stats.stackexchange.com/questions/275380/maximum-likelihood-estimation-for-bernoulli-distribution

Example I: Bernoulli

1. Write as joint distribution over all X

$$\prod_{i=1}^{n} p_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-xi}$$

2. Take logarithms

$$\log \prod_{i=1}^{n} p_\theta(x_i) = \sum_{i=1}^{n} x_i \log \theta + (1-x_i) \log(1-\theta)$$

$$= L(\theta)$$

3. Find derivative. Set $\sum_{i=1}^{n} x_i = \bar{X}$

$$\frac{d}{d\theta} L(\theta) = \frac{\bar{X}}{\theta} - \frac{n-\bar{X}}{1-\theta}$$

4. Set equal to zero and solve for $\hat{\theta}$:

$$\hat{\theta} = \frac{\bar{X}}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Derivation: https://stats.stackexchange.com/questions/275380/maximum-likelihood-estimation-for-bernoulli-distribution

Example II: Normal

- Know data comes from Normal distribution with variance 1 but unknown mean; find mean
- MLE for mean
  - $p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right)$
  - Maximising likelihood yields $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Exercise: derive MLE for *variance* σ^2 based on $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ with $\theta = (\mu, \sigma^2)$

Derivation: https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood

MLE ‘algorithm’

1. Given data $X_1, \ldots, X_n$ define probability distribution, p_θ, assumed to have generated the data
2. Express likelihood of data, $\prod_{i=1}^{n} p_\theta(X_i)$ (usually its *logarithm… why?)*
3. Optimise to find *best* (most likely) parameters $\hat{\theta}$
    1. take partial derivatives of log likelihood wrt θ
    2. set to 0 and solve (failing that, use gradient descent)

Decision theory

Wald

- Act to maximise utility - connected to economics and operations research
- Decision rule $\delta(\boldsymbol{x}) \in A$ an action space
  - ∗ E.g. Point estimate $\hat{\theta}(x_1, \dots, x_n)$
  - ∗ E.g. Out-of-sample prediction $\hat{Y}_{n+1}|X_1, Y_1, \dots, X_n, Y_n, X_{n+1}$
- Loss function l: economic cost, error metric
  - ∗ E.g. square loss of estimate $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
  - ∗ E.g. 0-1 loss of classifier predictions $l(y, \hat{y}) = 1[y \neq \hat{y}]$

COMP90051 Workshop 2 ¶

Bayesian inference

In this part of the workshop, we'll develop some intuition for priors and posteriors, which are crucial to Bayesian inference.

1. A lucky find

On the way to class, you discover an unusual coin on the ground.

As a dedicated student in statistical ML, you're interested in determining whether the coin is *biased*. More specifically, you want to estimate the probability θ that the coin will land heads-up when you toss it. If $\theta \approx \frac{1}{2}$ then we sy that the coin is *unbiased* (or fair).

You can use the function below to simulate a coin toss: it returns `1` for heads and `0` for tails.

2. Prior belief

Before you even toss the coin, you notice that the heads side appears to have more mass. Thus, your *prior belief* is that θ is slightly biased away from $\frac{1}{2}$ towards 0—i.e. you expect tails are more likely.

To quantify this prior belief, we assume that the prior distribution for θ is $\mathrm{Beta}(a, b)$, for some choice of the hyperparameters $a, b > 0$. (See link for info about the Beta distribution.) The prior probability density function for θ is therefore given by:

$$p(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

where $B(a, b)$ is a special function called the *Beta function*.

Select appropriate values for a and b by looking at the plot of $p(\theta)$ below.

3. Posterior updates

Now toss the coin once and denote the outcome by x_1.

We can update our belief about θ, based on this new evidence x_1. To do this we apply Bayes' rule to compute the posterior for θ:

$$p(\theta|x_1) = \frac{p(x_1|\theta)\, p(\theta)}{p(x_1)} \propto p(x_1|\theta)\, p(\theta)$$

where $p(\theta)$ is the prior given above and

$$p(x_1|\theta) = \theta^{x_1} (1-\theta)^{1-x_1}$$

is the likelihood.

Exercise: Show (on paper) that

$$p(\theta|x_1) \propto \theta^{x_1+a-1}(1-\theta)^{(1-x_1)+b-1}$$

which implies that $\theta|x_1 \sim \mathrm{Beta}[x_1+a,(1-x_1)+b]$.

Solution: Using Bayes' Theorem, we combine the Bernoulli likelihood with the Beta prior. We can drop constant factors and recover the normalising constants by comparing with the standard Beta distribution at the end.

$$\begin{aligned} p(\theta|x_1) &\propto p(x_1|\theta)\,p(\theta) \\ &\propto \theta^{\alpha+x_1-1}(1-\theta)^{\beta-x_1} \\ &\propto \frac{1}{B(\alpha',\beta')}\theta^{\alpha'-1}(1-\theta)^{\beta'-1} \end{aligned}$$

where $\alpha' = \alpha + x_1$ and $\beta' = \beta + 1 - x_1$.

Toss the coin a second time, denoting the outcome by x_2.

Again, we want to update our belief about θ based on the new information x_2. We take the previous posterior $p(\theta|x_1)$ as the new prior and apply Bayes' rule:

$$p(\theta|x_1,x_2) \propto p(x_2|\theta)p(\theta|x_1)$$

[Note: We assume the tosses are independent, otherwise the likelihood for x_2 would depend on x_1.] This gives $\theta|x_1,x_2 \sim \mathrm{Beta}[x_1+x_2+a,(2-x_1-x_2)+b]$.

Exercise: show that for n coin tosses, the posterior is $\theta|x_1,\ldots,x_n \sim \mathrm{Beta}[n_H+a,n-n_H+b]$ where $n_H = \sum_{i=1}^{n} x_i$ is the number of heads observed.

solution: We assume the coin tosses are i.i.d. with a probability θ of returning heads. The likelihood can be written as (where $\mathbf{x}_n = (x_1,\ldots x_n)$ is shorthand for all observations up to step n.):

$$p(\mathbf{x}_n|\theta) = \prod_{k=1}^{n} p(x_k|\theta) = \prod_{k=1}^{n} \theta^{x_k}(1-\theta)^{1-x_k}$$

Applying Bayes' theorem, the posterior assumes the form:

$$\begin{aligned} p(\theta|\mathbf{x}_n) &= p(\mathbf{x}_n|\theta)p(\theta) \\ &= p(\theta)\prod_{i=1}^{n} p(x_i|\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \\ &= \theta^{\sum_{i=1}^{n} x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^{n} x_i+\beta-1} \end{aligned}$$

This corresponds to $\mathrm{Beta}[n_H+a,n-n_H+b]$ by inspection.

4. MAP estimator and MLE estimator

The posterior $\theta|x_1,\ldots,x_n$ contains all the information we know about θ after observing n coin tosses. One way of obtaining a point estimate of θ from the posterior, is to take the value with the maximum a posteriori probability (MAP):

$$\begin{aligned} \hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta} p(\theta|x_1,\ldots,x_n) \\ &= \frac{n_H+a-1}{n+a+b-2} \end{aligned}$$

In general, the MAP estimator gives a different result to the maximum likelihood estimator (MLE) for θ:

Exercise: show that for n coin tosses, the posterior is $\theta|x_1,\dots,x_n \sim \mathrm{Beta}[n_H+a, n-n_H+b]$ where $n_H=\sum_{i=1}^n x_i$ is the number of heads observed.

solution: We assume the coin tosses are i.i.d. with a probability θ of returning heads. The likelihood can be written as (where $\mathbf{x}_n=(x_1,\dots x_n)$ is shorthand for all observations up to step n.):

$$p(\mathbf{x}_n|\theta)=\prod_{k=1}^n p(x_k|\theta)=\prod_{k=1}^n \theta^{x_k}(1-\theta)^{1-x_k}$$

Applying Bayes' theorem, the posterior assumes the form:

$$\begin{aligned}p(\theta|\mathbf{x}_n)&=p(\mathbf{x}_n|\theta)p(\theta)\\&=p(\theta)\prod_{i=1}^n p(x_i|\theta)\\&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\\&=\theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}\end{aligned}$$

This corresponds to $\mathrm{Beta}[n_H+a, n-n_H+b]$ by inspection.

---

4. MAP estimator and MLE estimator

The posterior $\theta|x_1,\dots,x_n$ contains all the information we know about θ after observing n coin tosses. One way of obtaining a point estimate of θ from the posterior, is to take the value with the maximum a posteriori probability (MAP):

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}}&=\arg\max_\theta p(\theta|x_1,\dots,x_n)\\&=\frac{n_H+a-1}{n+a+b-2}\end{aligned}$$

In general, the MAP estimator gives a different result to the maximum likelihood estimator (MLE) for θ:

$$\begin{aligned}\hat{\theta}_{\mathrm{MLE}}&=\arg\max_\theta p(x_1,\dots,x_n|\theta)\\&=\frac{n_H}{n}\end{aligned}$$

Exercise: How would you derive the above results for $\hat{\theta}_{\mathrm{MAP}}$ and $\hat{\theta}_{\mathrm{MLE}}$? Setup the equations necessry to solve for $\hat{\theta}_{\mathrm{MAP}}$ and $\hat{\theta}_{\mathrm{MLE}}$. You do not need to solve the equations at this stage.

Solution:

In a previous exercise, we found that the posterior was given by

$$p(\theta|x_1,\dots,x_n)\propto\theta^{n_H+a-1}(1-\theta)^{n-n_H+b-1}$$

The maximum a-posteriori estimate $\hat{\theta}_{\mathrm{MAP}}$ corresponds to the mode of this distribution. We can find the mode of a Beta pmf $f(\theta)\propto\theta^{\alpha-1}(1-\theta)^{\beta-1}$ by solving for the critical point $\tilde{\theta}$ as follows:

$$\frac{\partial f}{\partial\theta}(\tilde{\theta})\propto\tilde{\theta}^{\alpha-2}(1-\tilde{\theta})^{\beta-2}\left[(\alpha-1)-\tilde{\theta}(\alpha+\beta-2)\right]=0.$$

The solutions are $\tilde{\theta}=0, 1, \frac{\alpha-1}{\alpha+\beta-2}$ assuming $\alpha,\beta>1$. By performing the second derivative test, we find that $\frac{\alpha-1}{\alpha+\beta-2}$ corresponds to the maximum.

Thus we have

$$\hat{\theta}_{\mathrm{MAP}}=\frac{n_H+a-1}{n+a+b-2}$$

To find the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$, we undertake a similar procedure, replacing the posterior with the likelihood function:

To find the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$, we undertake a similar procedure, replacing the posterior with the likelihood function:

$$L(\theta) = p(x_1, \dots, x_n|\theta) = \theta^{n_H}(1-\theta)^{n-n_H}.$$

The stationary points $\tilde{\theta}$ of $L(\theta)$ stisfy:

$$\frac{\partial L}{\partial \theta}(\tilde{\theta}) = \tilde{\theta}^{n_H-1}(1-\tilde{\theta})^{n-n_H-1}\left[n_H(1-\tilde{\theta}) - (n-n_H)\tilde{\theta}\right] = 0$$

The solutions are $\tilde{\theta} = 0, 1, \frac{n_H}{n}$ assuming $n_H > 1$. The maximum occurs at $\frac{n_H}{n}$ which correspond to $\hat{\theta}_{\mathrm{MLE}}$.

Extension (Only return to this if you have completed the remaining workshop): Solve the equations you derived above. Give the condition for the estimators to be (exactly) equal, i.e. $\hat{\theta}_{\mathrm{MAP}} \equiv \hat{\theta}_{\mathrm{MLE}}$. What is the prior in this case?

$\hat{\Theta}$ is an estimate of fixed parameter Θ and no inherent noise or irreducible randomness associated with it, where randomness comes from solely the sampling variablity due to different training dataset. For $E_{X,Y}[(Y - \hat{f}(X))^2]$, where Y is the response variable that depends on certain underlying stochastic process with random noise. $\hat{f}(X)$ is a prediction function learned from data. No matter how well we learn the model we cannot eliminate the random noise.

5. Convergence of the estimates

Let's now toss the coin an additional 48 times (so that $n = 50$), recording $\hat{\theta}_{\mathrm{MLE}}$ and $\hat{\theta}_{\mathrm{MAP}}$ after each toss.

We plot the results below.

Questions:

1. Is the coin biased?
2. Do the MAP and MLE estimates converge to the sme value for θ?
3. What happens if you set $a = 1; b = 1$?
4. How does the posterior distribution for θ compare to the prior plotted above? (Use the code block below to plot the posterior.)

Solutions:

1. Yes
2. Yes
3. The MAP and MLE estimators are identical.
4. It's more concentrated.

Finally, we'll visualize the evolution of the posterior distribution as we observe more data. Before running the code cell below, take a couple of minutes to discuss with those around you how you expect the posterior to behave qualitatively) as the number of observed smples x_n increases.

Bonus material: Bayesian credible intervals

In principle, the posterior distribution contains all the information about the possible values of the parameter θ. To show the utility of the posterior, we can obtain a quantitative measure of the posterior uncertainty by computing a central (or equal-tailed) interval of posterior probability. These are known as *Bayesian credible intervals* and should not be confused with the frequentist concept of *confidence intervals* which leverage the distribution of point estimators. For a Bayesian credible interval, an e.g. 95% credible interval contains the true parameter value with 95% probability. In general, for a $1 - \alpha$ interval, where $\alpha \in (0, 1)$, this corresponds to the range of values $I = (\theta_1, \theta_2)$ above and below which lie exactly $\alpha/2$ of the posterior probability. That is, $\alpha/2$ of the probability mass of the posterior lies below θ_1, and $\alpha/2$ of the probability mass lies above θ_2.

We should check that $1 - \alpha$ of the probability mass actually lies inside our computed interval. That is, we expect

$$\int_{\theta_1}^{\theta_2} d\theta \; p(\theta|x_1, \dots x_n) = 1 - \alpha$$

Risk & Empirical Risk Minimisation (ERM)

- In decision theory, really care about *expected* loss
- Risk $R_\theta[\delta] = \mathrm{E}_{\boldsymbol{X}\sim\theta}[l(\delta(\boldsymbol{X}),\theta)]$ or

  $R_\theta[\delta] = \mathrm{E}_{(\boldsymbol{X},Y)\sim\theta}[l(\delta(\boldsymbol{X}),Y)]$
  - E.g. true test error, aka generalization error
- Want: Choose δ to minimise $R_\theta[\delta]$
- Can't directly! Why?
- ERM: Use training set ***X*** to approximate R_θ
  - Minimise empirical risk $\hat{R}_\theta[\delta] = \frac{1}{n}\sum_{i=1}^{n} l(\delta(X_i),\theta)$ or

    $\hat{R}_\theta[\delta] = \frac{1}{n}\sum_{i=1}^{n} l(\delta(X_i),Y_i)$

Decision theory vs. Bias-variance

We've already seen

- Bias: $\mathrm{B}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[\hat{\theta}(X_1, \dots, X_n)\right] - \theta$
- Variance: $\mathrm{Var}_\theta\left(\hat{\theta}\right) = \mathrm{E}_\theta\left[(\hat{\theta} - \mathrm{E}_\theta[\hat{\theta}])^2\right]$

But are they equally important? How related?

- Bias-variance decomposition of square-loss risk

$$\mathrm{E}_\theta\left[\left(\theta - \hat{\theta}\right)^2\right] = [\mathrm{B}(\hat{\theta})]^2 + \mathrm{Var}_\theta(\hat{\theta})$$

Tools of probabilistic inference

Bayes

- Bayesian probabilistic inference
  - Start with prior $\mathrm{P}(\theta)$ and likelihood $\mathrm{P}(X|\theta)$
  - Observe data $X = x$
  - Update prior to posterior $\mathrm{P}(\theta|X = x)$
- Primary tools to obtain the posterior
  - Bayes Rule: reverses order of conditioning

$$P(\theta|X = x) = \frac{P(X = x|\theta)P(\theta)}{P(X = x)}$$

  - Marginalisation: eliminates unwanted variables

$$\mathrm{P}(X = x) = \sum_t P(X = x, \theta = t)$$

These are general tools of probability and not specific to Bayesian stats/ML

This quantity is called the evidence

Example

- **Given:** The data comes from a **Normal distribution** with **variance 1** but **unknown mean** θ.
- **Goal:** Find the **posterior over the mean** after seeing one data point where **X=1**. Assume a **Normal prior** over θ with mean 0 and variance 1.

The Normal distribution:

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

$$P(\theta|X=1) = \frac{P(X=1|\theta)P(\theta)}{P(X=1)}$$

$$\propto P(X=1|\theta)P(\theta)$$

Discard constants wrt θ

$$= \left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{(1-\theta)^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi}} exp\left(-\frac{\theta^2}{2}\right)\right]$$

$$\propto exp\left(-\frac{(1-\theta)^2+\theta^2}{2}\right) = exp\left(-\frac{2\theta^2 - 2\theta + 1}{2}\right)$$

$$= exp\left(-\frac{\theta^2 - \theta + \frac{1}{2}}{2 \times \frac{1}{2}}\right)$$

Make leading numerator zero: $\times \frac{1}{2}$ on top and bottom

Example

$$P(\theta|X=1) = exp\left(-\frac{\theta^2 - \theta + \frac{1}{2}}{2\times\frac{1}{2}}\right)$$

$$= exp\left(-\frac{(\theta - \frac{1}{2})^2 + \frac{1}{4}}{2\times\frac{1}{2}}\right)$$

$$= exp\left(-\frac{(\theta - \frac{1}{2})^2}{2\times\frac{1}{2}}\right) .\ exp\left(-\frac{\frac{1}{4}}{2\times\frac{1}{2}}\right)$$

$$\propto exp\left(-\frac{(\theta - \frac{1}{2})^2}{2\times\frac{1}{2}}\right)$$

$$\propto \mathcal{N}(0.5,0.5)$$

The square completion trick!

https://en.wikipedia.org/wiki/Completing_the_square

$$\theta^2 - \theta + \frac{1}{2} \rightarrow (\theta - ?)^2$$

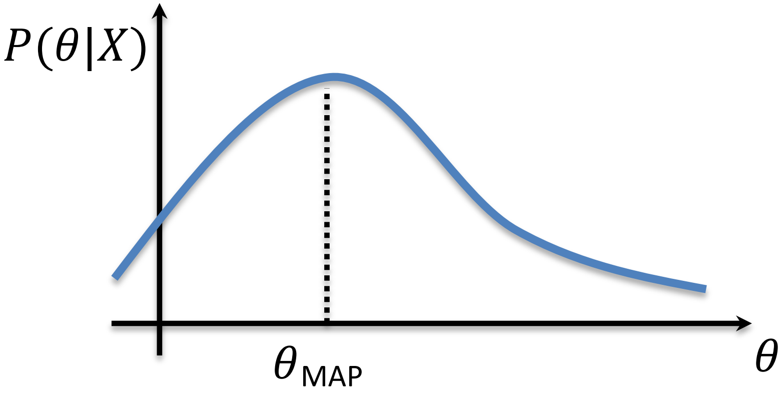
$$\theta^2 - \theta + \frac{1}{2} = (\theta - \frac{1}{2})^2 + \frac{1}{4}$$

The Normal distribution:

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

How Bayesians make point estimates

- They don't, unless forced at gunpoint!
  - The posterior carries full information, why discard it?
- But, there are common approaches
  - Posterior mean $E_{\theta|X}[\theta] = \int \theta P(\theta|X) d\theta$
  - Posterior mode $\underset{\theta}{\operatorname{argmax}} P(\theta|X)$ (max a posteriori or MAP)
  - There're Bayesian decision-theoretic interpretations of these



MLE in Bayesian context

- MLE formulation: find parameters that best fit data

$$\hat{\theta} \in \operatorname{argmax}_\theta P(X = x|\theta)$$

- Consider the MAP under a Bayesian formulation

$$\begin{aligned}\hat{\theta} &\in \operatorname{argmax}_\theta P(\theta|X = x) \\ &= \operatorname{argmax}_\theta \frac{P(X = x|\theta)P(\theta)}{P(X = x)} \\ &= \operatorname{argmax}_\theta P(X = x|\theta)P(\theta)\end{aligned}$$

- **Prior** $P(\theta)$ weights; MLE like *uniform* $P(\theta) \propto 1$