Probabilistic ML III

CS771: Introduction to Machine Learning

Purushottam Kar

Scaling in SGD

CSVM objective is $\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\left[1 - y^i\mathbf{w}^\top\mathbf{x}^i\right]_+$

If we had done GD, we would have summed up gradients w.r.t n data points (apart from gradient from regularizer)

In SGD, we get gradient from only one (random) point. To get same effect as in GD, popular to multiply that gradient by n

Similar to assuming that there are n clones of that data point for this update

In lec7-8.py, this was correctly done for the update for w but not for b

Corrected code uploaded onto GitHub – please pull

Error occurred since I first coded the solver with b hidden inside $\mathbf{w}$ – then I changed the code to make updates more explicit but forgot scaling for b ☹

Recap of Last Lecture

Bernoulli and Rademacher distributions over binary support

Categorical distributions over finite support with > 2 elements

Probabilistic Classfn: predict a PMF over all classes for each datapoint

Logistic Regression: map $\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x})$ so that $\mathbb{P}[y \mid \mathbf{x}^t, \mathbf{w}] = \sigma(y \cdot \mathbf{w}^\top \mathbf{x}^t)$

Likelihood function: probability assigned to true label i.e. $\mathbb{P}[y^t \mid \mathbf{x}^t, \mathbf{w}]$

Maximum Likelihood Estimate: the model that maximizes the likelihood function i.e. assigns "largest probability" to observed labels

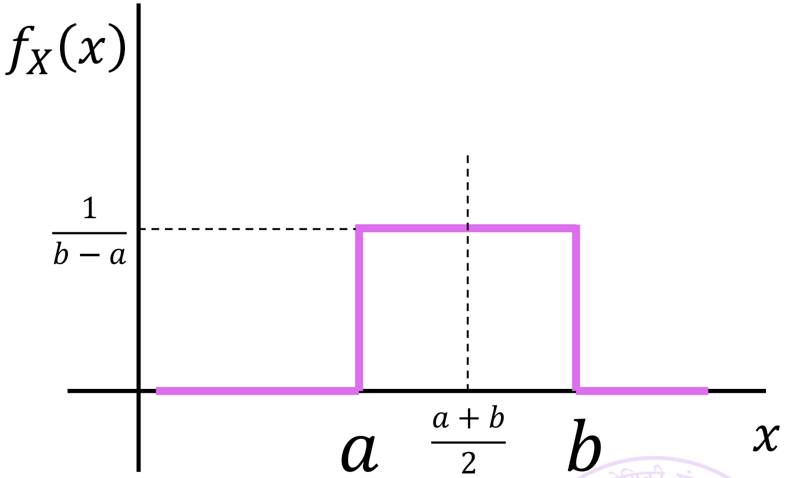
Softmax Regression: probabilistic multiclassification (cross entropy loss)

Continuous R.V.s: probability density function, rules revisited

Uniform Distribution

Can be defined over any finite interval

Let X be a continuous r.v. with support $S_X = [a, b] \in \mathbb{R}$. Then X is said to have a uniform distribution if its PDF is a constant function (uniform density) i.e. $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases}$



Note: $f_X(x) \geq 0$ if $x \in S_X$ and $\int_{S_X} f_X(t)\,dt = 1$

Mean: $\mathbb{E}[X] = (a + b)/2$

Variance: $\mathbb{V}[X] = (b - a)^2/12$

Note: variance increases as $b - a \uparrow$ since r.v. more "spread out"

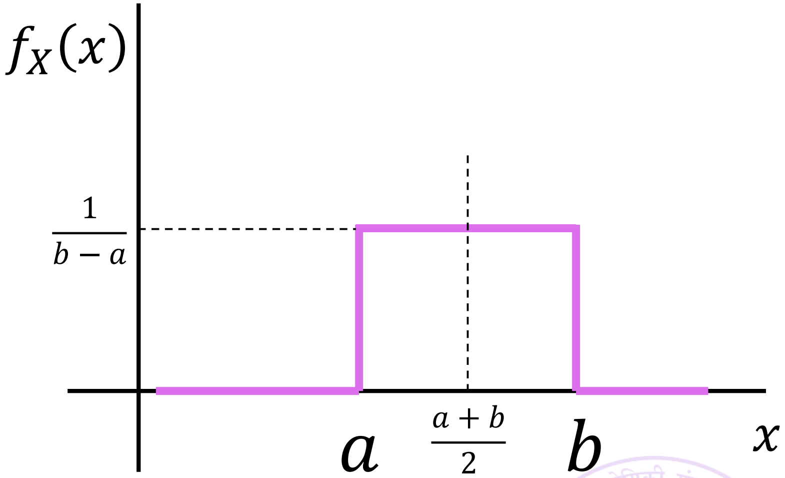
Notation: Often we use $\text{UNIF}([a, b])$ to denote uniform dist over $[a, b]$

Uniform Distribution

Recall that we commented that although we must have $f_X(x) > 0$, we need not have $f_X(x) \leq 1$. Note that if in the uniform case, if we have $b - a < 1$ then indeed $f_X(x) > 1$ and its perfectly fine

Can be d

Let X be a continuous r.v. with support $S_X = [a, b] \in \mathbb{R}$. Then to have a uniform distribution if its PDF is a constant function (uniform density) i.e. $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases}$



Note: $f_X(x) \geq 0$ if $x \in S_X$ and $\int_{S_X} f_X(t)\, dt = 1$

Mean: $\mathbb{E}[X] = (a + b)/2$

Variance: $\mathbb{V}[X] = (b - a)^2/12$

Note: variance increases as $b - a \uparrow$ since r.v. more "spread out"

Notation: Often we use $\text{UNIF}([a, b])$ to denote uniform dist over $[a, b]$

Gaussian (aka Normal) Distributions

Arguably one of the most popular of all probability distributions

Models our intuitive assumption that in real life, data often takes values around its mean value and it gets unlikely to witness extreme values

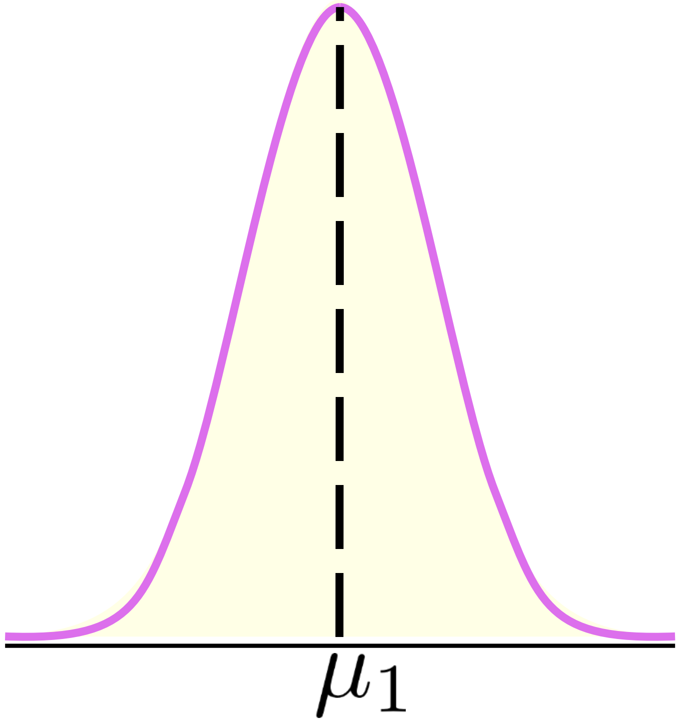
- *A fundamental result in probability theory* – the law of large numbers –*shows that some form of this is indeed true*

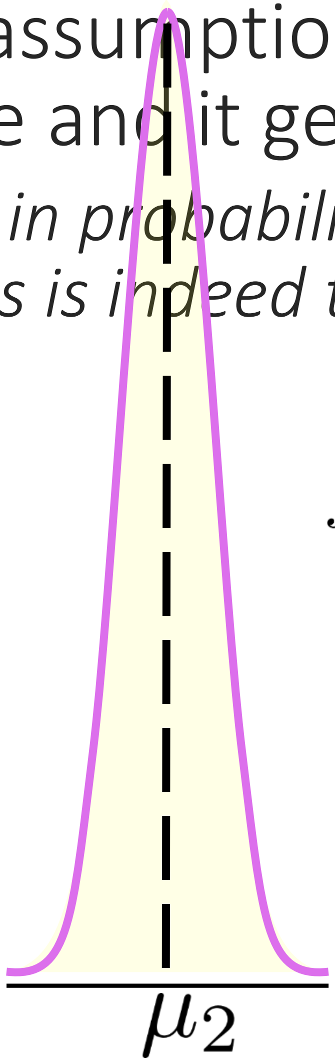
Gaussian (aka Normal) Distributions

Arguably one of the most popular of all probability distributions

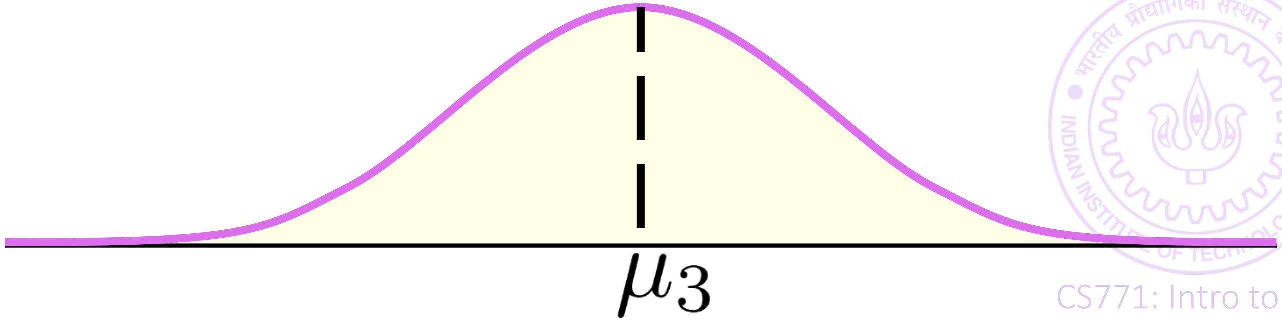
Models our intuitive assumption that in real life, data often takes values around its mean value and it gets unlikely to witness extreme values

A fundamental result in probability theory – the law of large numbers –*shows that some form of this is indeed true*





$$f_X[x \mid \mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Gaussian Distributions

Specifying a Bernoulli/Rademacher distribution took two numbers,

Specifying a categorical distribution over C elements takes C numbers

Specifying a Gaussian distribution over $\mathbb{R}$ requires two numbers

μ: must be a real number (may be negative or positive or even zero)

σ^2: must be a non-negative real number

Notation: PDF for a Gaussian r.v. X i.e. $f_X[X \mid \mu, \sigma^2]$ is often written as $\mathcal{N}_X(x; \mu, \sigma^2)$ or simply as $\mathcal{N}(x; \mu, \sigma^2)$

Notice that even here we condition on constants (either using | or ; symbol)

The notation is no accident – if the PDF of a r.v. X is $\mathcal{N}_X(x; \mu, \sigma^2)$, then

$\mathbb{E}[X] = \mu$ *= Mode = Median, as well as* $\mathbb{V}[X] = \sigma^2$

Requires a bit of integration to prove these results ☺

Gaussian

Specifying a Be r

Specifying a cat umbers

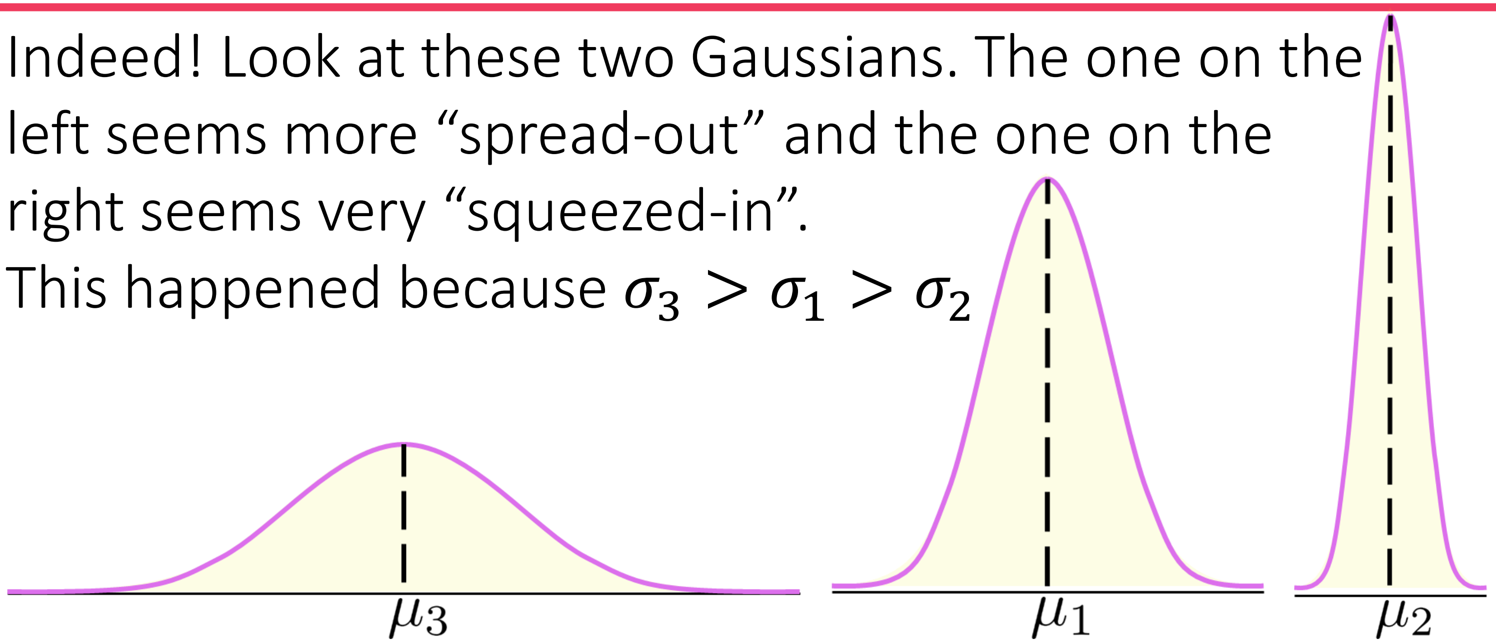
Specifying a Ga ers

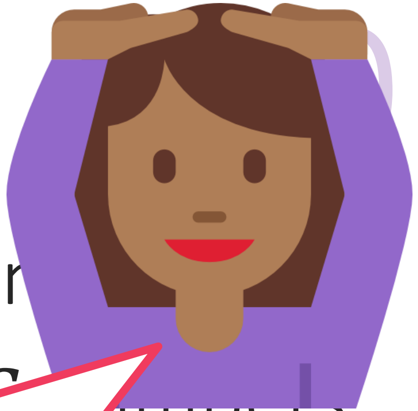
μ: must be a r o)

σ^2: must be a non-negative real number

Indeed! Look at these two Gaussians. The one on the left seems more “spread-out” and the one on the right seems very “squeezed-in”.

This happened because $\sigma_3 > \sigma_1 > \sigma_2$





Notation: PDF for a Gaussian r.v. X i.e. $f_X[X \mid \mu, \sigma^2]$ is often written as $\mathcal{N}_X(x; \mu, \sigma^2)$ or simply as $\mathcal{N}(x; \mu, \sigma^2)$

Notice that even here we condition on constants (either using | or ; symbol)

The notation is no accident – if the PDF of a r.v. X is $\mathcal{N}_X(x; \mu, \sigma^2)$, then

$\mathbb{E}[X] = \mu$ = *Mode = Median, as well as* $\mathbb{V}[X] = \sigma^2$

Requires a bit of integration to prove these results ☺

Operations with Gaussians

Since integration can be a pain, some handy results about Gaussians

Let X, Y be two **independent** r.v. whose PDF is Gaussian i.e. $\mathcal{N}_X(\cdot\,; \mu_X, \sigma_X^2)$ and $\mathcal{N}_Y(\cdot\,; \mu_Y, \sigma_Y^2)$. Then we have

Scaling Rule: If $Z = c \cdot X$ then Z is also Gaussian $\mathcal{N}_Z(\cdot\,; c \cdot \mu_X, c^2 \cdot \sigma_X^2)$

Sum Rule: If $W = X + Y$ then W is also Gaussian too

$$\mathcal{N}_W(\cdot\,; \mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Shift Rule: If $V = X + c$ (c const) then V is Gaussian $\mathcal{N}_V(\cdot\,; \mu_X + c, \sigma_X^2)$

Tail Rule: $\{\mathbb{P}[X \geq \mu_X + t \cdot \sigma_X] = \mathbb{P}[X \leq \mu_X - t \cdot \sigma_X]\} \leq e^{-t^2/2}$

It gets exponentially less likely that a Gaussian r.v. takes value far from mean

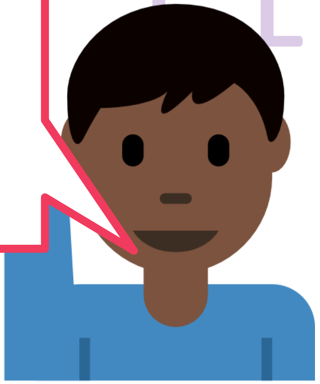
For $t = 5$, we have $\mathbb{P}[|X - \mu_X| \geq 5 \cdot \sigma_X] < 0.000004$ (5-sigma rule)

As $\sigma_X \downarrow$ the r.v. gets more and more concentrated around its mean

O

Si

Note: we can derive results such as $\mathbb{E}W = \mu_X + \mu_Y$ and $\mathbb{V}W = \sigma_X^2 + \sigma_Y^2$ using rules we studied earlier. However, those rules do not assure us that W must be Gaussian (they just assure us that W is some r.v. with such and such mean and variance. It takes special analysis to show that Z, W, V etc are Gaussian r.v. too!



Let X, Y be two **independent** r.v. whose PDF is Gaussian i.e. $\mathcal{N}_X(\cdot\,; \mu_X, \sigma_X^2)$ and $\mathcal{N}_Y(\cdot\,; \mu_Y, \sigma_Y^2)$. Then we have

Scaling Rule: If $Z = c \cdot X$ then Z is also Gaussian $\mathcal{N}_Z(\cdot\,; c \cdot \mu_X, c^2 \cdot \sigma_X^2)$

Sum Rule: If $W = X + Y$ then W is also Gaussian too

$$\mathcal{N}_W(\cdot\,; \mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Shift Rule: If $V = X + c$ (c const) then V is Gaussian $\mathcal{N}_V(\cdot\,; \mu_X + c, \sigma_X^2)$

Tail Rule: $\{\mathbb{P}[X \geq \mu_X + t \cdot \sigma_X] = \mathbb{P}[X \leq \mu_X - t \cdot \sigma_X]\} \leq e^{-t^2/2}$

It gets exponentially less likely that a Gaussian r.v. takes value far from mean

For $t = 5$, we have $\mathbb{P}[|X - \mu_X| \geq 5 \cdot \sigma_X] < 0.000004$ (5-sigma rule)

As $\sigma_X \downarrow$ the r.v. gets more and more concentrated around its mean

Laplacian Distribution

Close cousins of Gaussian distributions except that a Laplacian r.v. concentrates much more strongly around its mean than a Gaussian r.v.

Also require two parameters to be specified $\mu \in \mathbb{R}, \sigma \geq 0$

Mean = Mode = Median: μ, **Variance:** $2\sigma^2$

Laplacian Distribution

If X is a r.v. with a Laplacian PDF with parameters μ_X, σ_X, then $Y = a \cdot X + b$ (where a, b are constants) is also a Laplacian r.v. but with parameters $\mu_Y = a \cdot \mu_X + b$ and $\sigma_Y = a \cdot \sigma_Y$

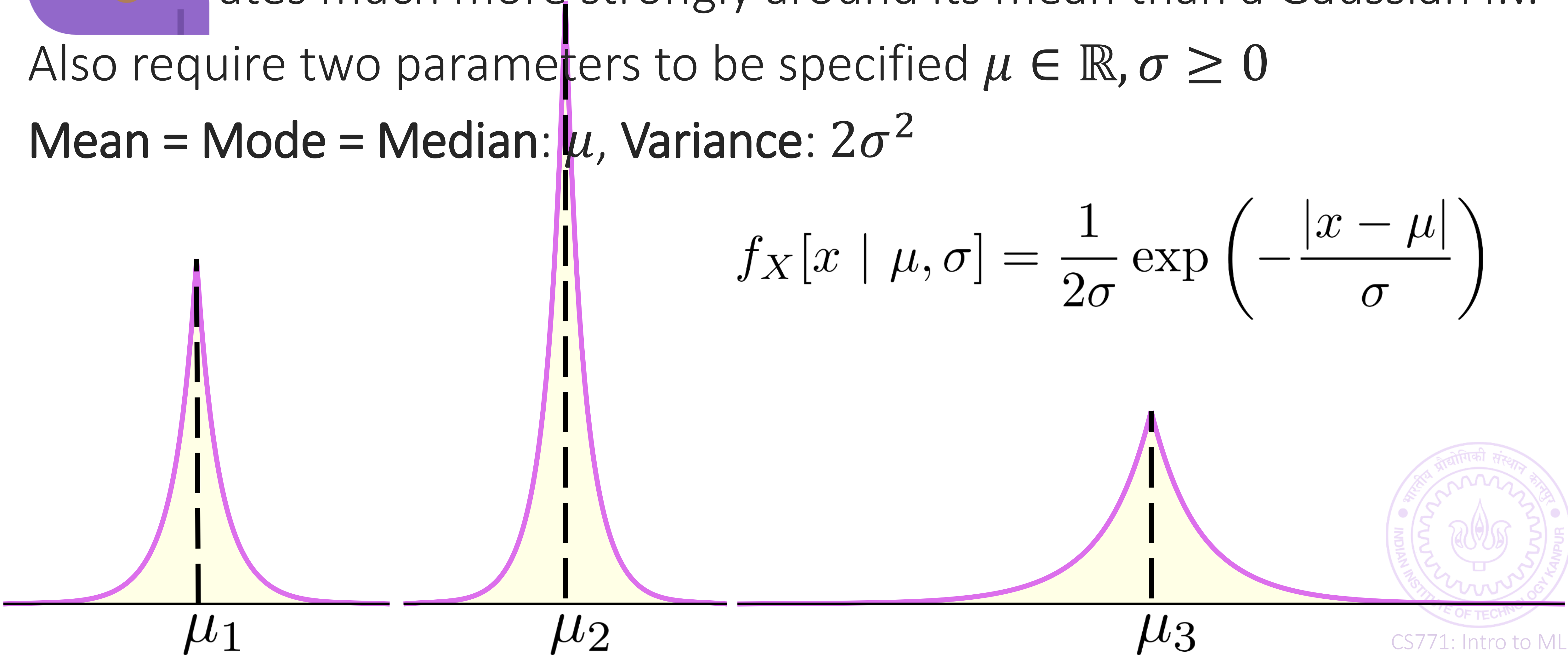
acian r.v.

rates much more strongly around its mean than a Gaussian r.v.

Also require two parameters to be specified $\mu \in \mathbb{R}, \sigma \geq 0$

Mean = Mode = Median: μ, Variance: $2\sigma^2$

$$f_X[x \mid \mu, \sigma] = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right)$$



Probabilistic Regression

In order to perform probabilistic regression I have to assign a label distribution over all $\mathbb{R}$ for every data point $\mathbf{x}$

Suppose I decide to do that using a Gaussian distribution – need to decide on a mean $\mu_{\mathbf{x}}$ and a variance $\sigma_{\mathbf{x}}^2 > 0$

Popular choice: Let $\mu_{\mathbf{x}} = \mathbf{w}^\top \mathbf{x}$ and $\sigma_{\mathbf{x}}^2 = \sigma^2$ i.e. $\mathcal{N}(\cdot \mid \mathbf{w}^\top \mathbf{x}, \sigma^2)$

We can also choose a different σ for every data point – more complicated

Likelihood function w.r.t a data point (x^i, y^i) then becomes

$$\mathcal{N}(y^i \mid \mathbf{w}^\top \mathbf{x}^i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\left(y^i - \mathbf{w}^\top \mathbf{x}^i\right)^2 / 2\sigma^2\right)$$

Negative log likelihood w.r.t a set of data points $\left\{(x^i, y^i)\right\}_{i=1}^n$

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{n}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n \left(y^i - \mathbf{w}^\top \mathbf{x}^i\right)^2$$

Probabilistic Regression

In order to perform probabilistic regression I have to assign a label distribution over all $\mathbb{R}$ for every data point $\mathbf{x}$

Suppose I decide to do that using a Gaussian distribution – need to decide on a mean $\mu_{\mathbf{x}}$ and a variance $\sigma_{\mathbf{x}}^2 > 0$

Popular choice: Let $\mu_{\mathbf{x}} = \mathbf{w}^\top\mathbf{x}$ and $\sigma_{\mathbf{x}}^2 = \sigma^2$ i.e. $\mathcal{N}(\cdot \mid \mathbf{w}^\top\mathbf{x}, \sigma^2)$

We can also choose a different σ for every data point – more complicated

Likelihood function w.r.t a data point (x^i, y^i) then becomes

$$\mathcal{N}(y^i \mid \mathbf{w}^\top\mathbf{x}^i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\left(y^i - \mathbf{w}^\top\mathbf{x}^i\right)^2 / 2\sigma^2\right)$$

Negative log likelihood w.r.t a set of data points $\{(x^i, y^i)\}_{i=1}^n$

$$\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^n \left(y^i - \mathbf{w}^\top\mathbf{x}^i\right)^2$$

Probabilistic R

In order to perform a

distribution over all ℝ for every data point $\mathbf{x}$

Suppose I decide to do that using a G eed to

decide on a mean $\mu_{\mathbf{x}}$ and a variance

Popular choice: Let $\mu_{\mathbf{x}} = \mathbf{w}^\top\mathbf{x}$ and $\sigma_{\mathbf{x}}^2 = \sigma^2$ i.e. $\mathcal{N}(\cdot \mid \mathbf{w}^\top\mathbf{x}, \sigma^2)$

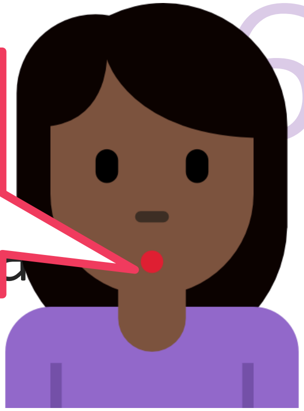
We can also choose a different σ fo cated

Likelihood function w.r.t a data po

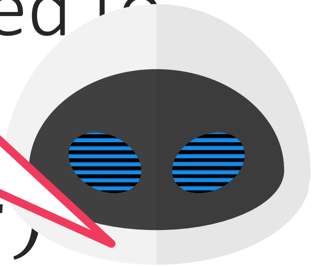
$$\mathcal{N}(y^i \mid \mathbf{w}^\top\mathbf{x}^i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\left(y^i - \mathbf{w}^\top\mathbf{x}^i\right)^2 / 2\sigma^2\right)$$

Negative log likelihood w.r.t a set of data points $\{(x^i, y^i)\}_{i=1}^n$

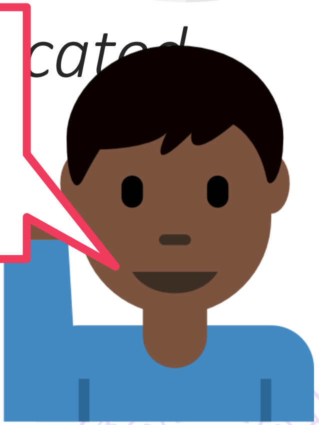
$$\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^n \left(y^i - \mathbf{w}^\top\mathbf{x}^i\right)^2$$



But apart from the first term and the scaling factor, both of which are constants and do not depend on the model $\mathbf{w}$ the rest is just the least squares loss term!



The MLE with respect to the Gaussian likelihood indeed the minimizes least squares loss



Also note that if we set all $\sigma_{\mathbf{x}}^2 = \sigma^2$ then it does not matter which σ we choose – will get the same model

Probabilistic Regression

Suppose I decide to use a Laplacian distribution instead and choose $\mu_{\mathbf{x}} = \mathbf{w}^\top\mathbf{x}$ and $\sigma_{\mathbf{x}} = \sigma$ i.e. $\mathrm{Lap}(\cdot \mid \mathbf{w}^\top\mathbf{x}, \sigma^2)$

Likelihood function w.r.t a data point (x^i, y^i) then becomes $\mathrm{Lap}(y^i \mid \mathbf{w}^\top\mathbf{x}^i, \sigma^2) = \frac{1}{2\sigma}\exp(-|y^i - \mathbf{w}^\top\mathbf{x}|/\sigma)$

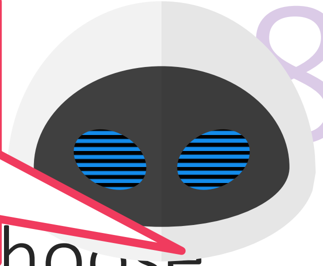
Negative log likelihood w.r.t a set of data points $\{(x^i, y^i)\}_{i=1}^n$

$$\min_{\mathbf{w}\in\mathbb{R}^d} n \cdot \ln(\sigma) + \frac{1}{\sigma}\sum_{i=1}^n |y^i - \mathbf{w}^\top\mathbf{x}^i| = \min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^n |y^i - \mathbf{w}^\top\mathbf{x}^i|$$

Thus, if we change the likelihood function to use the Laplacian distribution instead, the MLE ends up minimizing absolute loss!

As before, does not matter which σ we choose

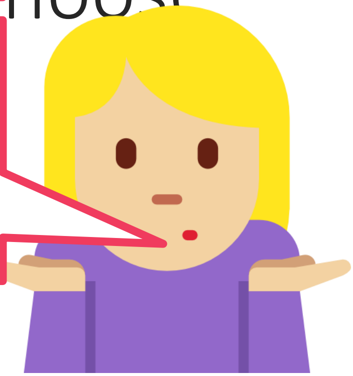
Proba



Be warned though – the σ we chose will start mattering the moment we add regularization! It is just that in these simple cases it does not matter. σ is usually treated like a hyperparameter and tuned.

Suppose I decide to use a Laplacian distribution instead and choose

$\mu_{\mathbf{x}} = \mathbf{w}$

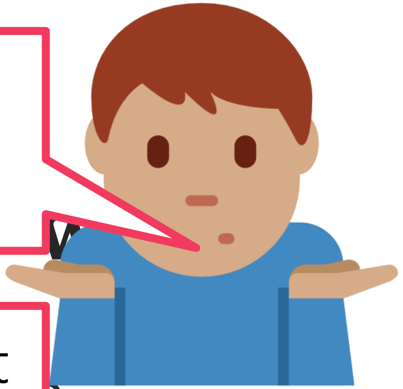


So I am a bit confused. All MLEs (classification/regression) demand a model that places maximum probability on the true label. Why don't we just ask the model to predict the true label itself?

Likelihoo

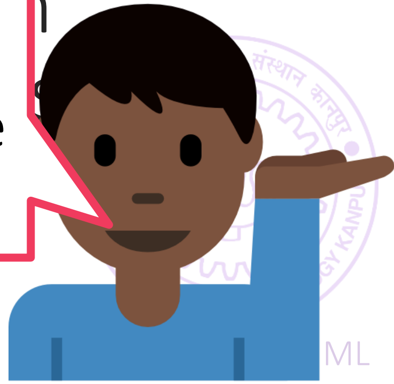
$\mathrm{Lap}\left(y^i \mid \mathbf{w}^\top \mathbf{x}^i, \sigma^2\right) = \frac{1}{2\sigma} \exp\left(-\left|y^i - \mathbf{w}^\top \mathbf{x}\right| / \sigma\right)$

Negative



That is like asking the PMF/PDF to place probability 1 on the true label and 0 everywhere else – why can't we do just that?

$\min_{\mathbf{w} \in \mathbb{R}^d} n$



For the same reason we needed slack variables in CSVM – to allow for the fact that in realistic situations, no linear model may be able to do what we would ideally like. In probabilistic ML, allowing the model to place a less than 1 probability on the true label is much like a slack – allows us to learn good models even if not perfect ones

Probabilistic Regularization??

We have seen that MLE often reduces to loss minimization e.g. logistic regression/least squares regression but without regularization terms ☹

Even probabilistic methods can do regularization by way of *priors*

Recall: regularization basically tells us which kinds of models we prefer

L2 regularization means we prefer models with small L2 norm

L1 regularization means we prefer models with small L1 norm/sparse models

In the language of probability, the most direct way of specifying such a preference is by specifying a probability distribution itself

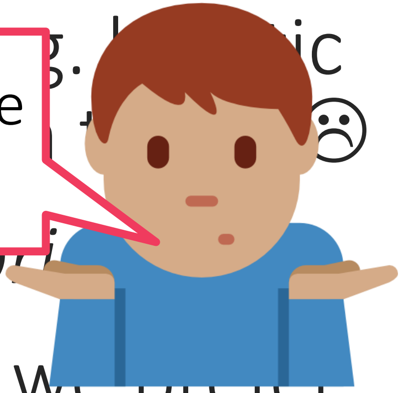
Prior: a probability distribution over all possible models

Just like we usually decide regularization before seeing any data, prior distribution also does not consider/condition on, any data

Probabilistic Regularization??

We have seen that MLE o g. ic regression/least squares r ☹

Even probabilistic methods can do regularization by way of pr

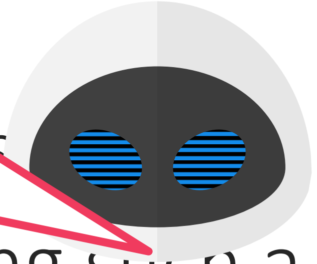


But our models are vectors right? Can we have probability distribution over vectors as well?

Recall: regularization basically tells us which kinds of models we prefer

L2 regularization means w

L1 regularization means w



Of course we can. But first, let us see the basic operations in a toy 1D setting before getting into the complications of vector-valued r.v.s

In the language of probability, the most direct way of specifying such a preference is by specifying a probability distribution itself

Prior: a probability distribution over all possible models

Just like we usually decide regularization before seeing any data, prior distribution also does not consider/condition on, any data

Can you Guess the Mean?

There is a Gaussian with unknown mean but known variance (for sake of simplicity) from which we receive n independent samples

$$x_1, x_2, \dots, x_n \sim \mathcal{N}(\mu^*, 1)$$

Can we estimate the “model” μ^* from these samples?

Likelihood function: for a candidate model μ and sample x_i

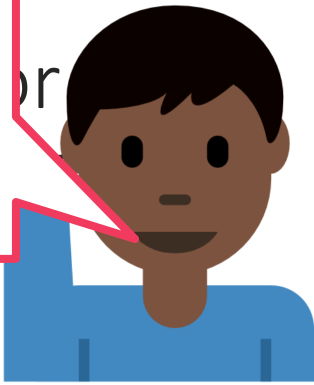
$$\mathbb{P}[x_i \mid \mu, 1] = \frac{1}{\sqrt{2\pi}} \exp(-(x_i - \mu)^2/2)$$

MLE: $\arg\max_{\mu\in\mathbb{R}} \prod_{i=1}^n \mathbb{P}[x_i \mid \mu, 1] = \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^n (x_i - \mu)^2$

Suppose we believe (e.g. someone tells us) even before the samples have been presented that μ^* definitely lies in the interval $[0,2]$ (but could otherwise be any value within that interval)

Can you Guess the Mean?

In this case we are said to have a *prior belief* or simply *prior*, on the models μ, in this case the uniform prior $\mathrm{UNIF}([0,2])$. This means that unless we see any data to make us believe otherwise, we will think $\mathbb{P}[\mu] = \begin{cases} 0.5 & \text{if } x \in [0,2] \\ 0 & \text{if } x \notin [0,2] \end{cases}$.



r

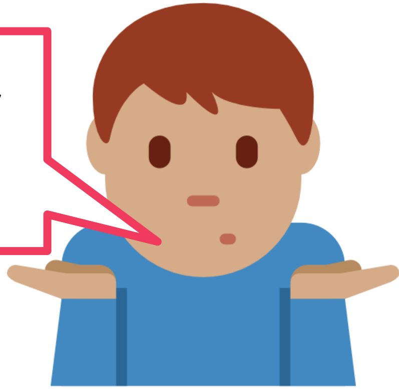
$x_1, x_2, \ldots, x_n \sim \mathcal{N}(\mu^*, 1)$

Can we estimate the “model” μ^* from these samples?

Likelihood function: for a c

$\mathbb{P}[x_i \mid \mu, 1] = \frac{1}{\sqrt{2\pi}}\exp(-($

What happens we do see some data, namely the actual samples from the distribution?



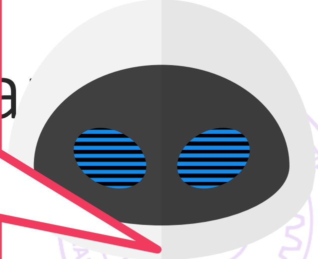
MLE: $\arg\max_{\mu\in\mathbb{R}} \prod_{i=1}^{n} \mathbb{P}[x_i \mid \mu, 1] = \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (x_i - \mu)^2$

Suppose we believe (e.g a

have been presented tha

could otherwise be any value within that interval)

We use the samples and the rules of probability to update our beliefs about what μ can and cannot be. Let us see how to do this



Posterior

Before we see any data, we have a *prior* belief $\mathbb{P}[\mu]$ on the models

It tells us which models are more likely/less likely before we have seen data

Then we see data $x_1, \ldots x_n$ and we wish to update our belief. Basically we want to find out $\mathbb{P}[\mu \mid x_1, \ldots, x_n]$

This quantity has a name: posterior belief *or simply* posterior

It tells us which models are more likely/less likely after we have seen data

$$\mathbb{P}[\mu \mid x_1, \ldots, x_n] = \frac{\mathbb{P}[x_1, \ldots, x_n \mid \mu] \cdot \mathbb{P}[\mu]}{\mathbb{P}[x_1, \ldots, x_n]} = \frac{\mathbb{P}[\mu] \cdot \prod_{i=1}^n \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^n \mathbb{P}[x_i]}$$

$$= \frac{\mathbb{P}[\mu] \cdot \prod_{i=1}^n \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^n \int_{\mathbb{R}} \mathbb{P}[x_i \mid t] \cdot \mathbb{P}[t] \, dt} = \frac{0.5 \cdot \prod_{i=1}^n \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^n 0.5 \cdot \int_0^2 \mathbb{P}[x_i \mid t] \, dt} \text{ if } \mu \in [0,2]$$

else if $\mu \notin [0,2]$, *then* $\mathbb{P}[\mu \mid x_1, \ldots, x_n] = 0$

Posterior

Before we see any data, we have a *prior* belief $\mathbb{P}[\mu]$ on the models

It tells us which models a … *en*

Then we see data $x_1, \ldots, x$ … we want to find out $\mathbb{P}[\mu$ … $_1, \ldots, x_n]$

T… has a name: … *ly* posterior

It tells us which models are more likely/less likely after we have seen data

$$\mathbb{P}[\mu \mid x_1, \ldots, x_n] = \frac{\mathbb{P}[x_1, \ldots, x_n \mid \mu] \cdot \mathbb{P}[\mu]}{\ldots} = \frac{\mathbb{P}[\mu] \cdot \prod_{i=1}^{n} \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^{n} \ldots}$$

$$= \frac{\mathbb{P}[\mu] \cdot \prod_{i=1}^{n} \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^{n} \int_{\mathbb{R}} \mathbb{P}[x_i \mid t] \cdot \mathbb{P}[t]\, dt} = \frac{0.5 \cdot \prod_{i=1}^{n} \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^{n} 0.5 \cdot \int_{0}^{2} \mathbb{P}[x_i \mid t]\, dt} \text{ if } \mu \in [0,2]$$

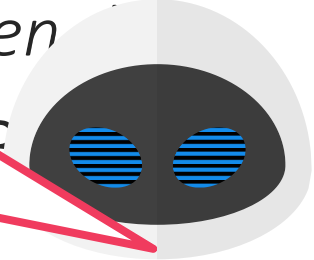
else if $\mu \notin [0,2]$*, then* $\mathbb{P}[\mu \mid x_1, \ldots, x_n] = 0$

Bayes Rule

Samples are independent

Law of total probability

$\mathbb{P}[\mu] = \mathrm{UNIF}([0,2])$



Keep in mind that when we say $\mathbb{P}[\mu]$ or $\mathbb{P}[\mu \mid x_1, \ldots, x_n]$, we mean probability density and not probability since μ is a continuous r.v.

Maximum a Posteriori (MAP) Estimate

Just as MLE gave us the model $\arg\max_{\mu \in \mathbb{R}} \mathbb{P}[x_1, \ldots, x_n \mid \mu, 1]$, MAP gives us the model $\arg\max_{\mu \in \mathbb{R}} \mathbb{P}[\mu \mid x_1, \ldots, x_n, 1] = \arg\max_{\mu \in \mathbb{R}} \frac{\mathbb{P}[\mu] \cdot \prod_{i=1}^{n} \mathbb{P}[x_i \mid \mu]}{\prod_{i=1}^{n} 0.5 \cdot \int_0^2 \mathbb{P}[x_i \mid t]\, dt}$

Thus, MAP returns the model that becomes the most likely one *after* we have seen some data

__Note__: posterior probability (density) of some models may be larger than their prior probability (density) i.e. after seeing data those models seem more likely, for other models, it may go down i.e. they seem less likely after seeing the data

__Note__: However, if prior probability (density) of some model is 0, the posterior probability (density) has to be zero as well – need to be careful about priors

__Warning__: Do not read too much into these names likelihood, prior, posterior. *All of them tell us how likely something is, given or not given something else*

Maximum a Posteriori (MAP) Estimate

Just as MLE gave us the mode

us the model $\arg\max_{\mu\in\mathbb{R}} \mathbb{P}[\mu \mid x_1, \ldots, x_n, 1] = \arg\max_{\mu\in\mathbb{R}} \frac{\mathbb{P}[\mu]\cdot\prod_{i=1}^{n} \quad \mu]}{\prod_{i=1}^{n} 0.5\cdot\int_0 \quad dt}$

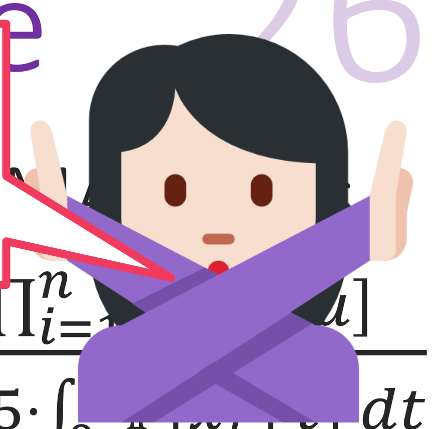
Thus, MAP returns the one after we have seen some da

Note: posterior probability (density) of some models may be larger t prior probability (de i.e. after seeing data those models seem more likely, for other models, it in data

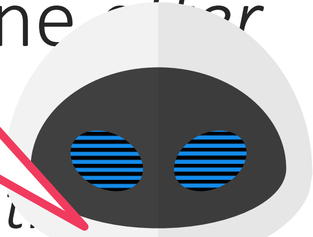
Note: However, if pr probability (density) ut

Warning: Do not read too much into these names likelihood, prior, p *All of them tell us how likely something is, given or not given something else*

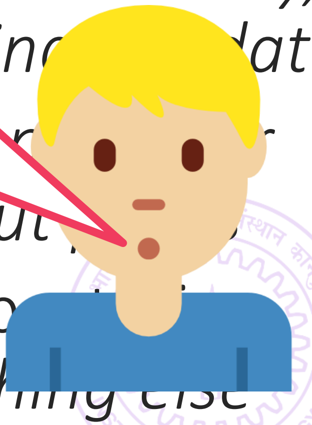
It is better to choose priors that do not completely exclude some models by giving them 0 probability (as we did)



True! Even in general, if your priors are bad, or too strong, then you may end up getting funny models as a result of doing MAP estimation



Indeed! For example if we were wrong and μ^* was actually not $\in [0,2]$ then not matter how many samples we see, we will never estimate μ^* correctly!!



MAP vs Regularization

$$\arg\max_{\mu\in\mathbb{R}} \frac{\mathbb{P}[x_1,\ldots,x_n \mid \mu]\cdot\mathbb{P}[\mu]}{\mathbb{P}[x_1,\ldots,x_n]} = \arg\max_{\mu\in\mathbb{R}} \mathbb{P}[x_1,\ldots,x_n \mid \mu]\cdot\mathbb{P}[\mu]$$

Taking negative log likelihoods on both sides $\mathbb{P}[\mu]\cdot\prod_{i=1}^{n}\mathbb{P}[x_i \mid \mu]$

$$\arg\min_{\mu\in\mathbb{R}} -\ln\mathbb{P}[\mu] + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

However, $\mathbb{P}[\mu]$ is constant for $\mu\in[0,2]$ and 0 otherwise ($\ln 0 \to \infty$)

$$\arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n}(x_i-\mu)^2 \text{ s.t. } \mu\in[0,2]$$

Thus, even MAP solutions can correspond to optimization problems!

In this case, what was the prior became a constraint

In general, the prior becomes a regularizer

MAP vs Regularization

Consider the same problem as before but a different prior

This time we do not believe μ must have been in the interval $[0,2]$ *but a much milder prior that μ is not too large*

A good way to express this is to use a Gaussian prior

$$\mathbb{P}[\mu] = \mathcal{N}(\mu\,;\,0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

MAP: $\arg\min_{\mu\in\mathbb{R}} -\ln\mathbb{P}[\mu] + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$

$$= \arg\min_{\mu\in\mathbb{R}} \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 = \arg\min_{\mu\in\mathbb{R}} \frac{\mu^2}{\sigma^2} + \sum_{i=1}^{n}(x_i-\mu)^2$$

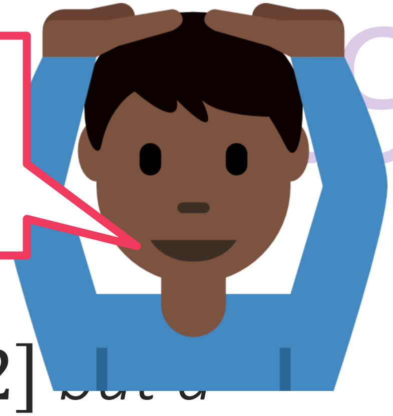
Thus, a Gaussian prior gave us L2 regularization!

Note: *σ effecitely dictates the regularization constant – not useless!!*

Note: *this is basically ridge regression except in one dimension!!*

MAP vs Regula

Similarly, had we used a Laplacian prior, we would have obtained L1 regularization instead



Consider the same problem as before but a different prior

This time we do not believe μ must have been in the interval $[0,2]$ *but a much milder prior that μ is not too large*

A good way to express this is to use a Gaussian prior

$$\mathbb{P}[\mu] = \mathcal{N}(\mu\,;\,0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

MAP: $\arg\min_{\mu\in\mathbb{R}} -\ln\mathbb{P}[\mu] + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$

$$= \arg\min_{\mu\in\mathbb{R}} \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 = \arg\min_{\mu\in\mathbb{R}} \frac{\mu^2}{\sigma^2} + \sum_{i=1}^{n}(x_i-\mu)^2$$

Thus, a Gaussian prior gave us L2 regularization!

***Note**: σ effecitely dictates the regularization constant – not useless!!*

***Note**: this is basically ridge regression except in one dimension!!*