Intro to Probabilistic Modelling in Machine Learning

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Material Covered

Models: Linear models, basis expansion, neural networks

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity

Settings: Regression, Classification

This lecture: Density estimation, Maximum likelihood estimation, maximum a posteriori estimation, classification, classification, logistic regression

Material Covered

Models: Linear models, basis expansion, neural networks

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity

Settings: Regression, Classification

Next lecture we will do a more detailed study on how to manipulate probability distributions

So far, we have mostly ignored probability

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}] = \frac{1}{N}\sum_{i=1}^{N}(\theta^{\top}x^{(i)} - y^{(i)})^2$$

So far, we have swept this under the rug as being just some unknown empirical distribution. Today, we will begin to establish how probabilistic treatment gives us different modelling choices and leads to different parameter estimates!

Starting from square one

Though we have focused on supervised learning settings, to introduce a probabilistic perspective, we strip away our labels and focus on density estimation

$x \in \mathbb{R}^n$ Feature vector/Input Space

$\{x_1, x_2, \ldots, x_N\}$ Dataset. Assump:, N >> n, iid

Maximum likelihood estimation

$x \in \mathbb{R}^n$ Feature vector/Input Space

$\{x_1, x_2, \ldots, x_N\}$ Dataset. Assump:, N >> n, iid

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution:

$$x \sim p_\theta$$

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution:

$$x \sim p_\theta$$

The exact form of the parameter we are interested is indexed by this parameter vector theta (as before)

$$\theta \in \Theta$$

$\theta \in \Theta$

We write the space this parameter lives in as big/capital Theta

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution

2) Write out the probability of observing our data from this distribution:

Recall we model our data/observations as random vars:

$$\{x_1 \sim X_1, x_2 \sim X_2, \ldots, x_N \sim X_N\}$$

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.
1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution:

Probability of the ith observation:

$$p(X_i = x_i|\theta)$$

prob of getting that outcome given params

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution

2) Write out the probability of observing our data from this distribution:

Recall that by independence:

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N|\theta)$$

$= P(X_1 = x_1) \& P(X_2 = x_2) \& \cdots$

$P(A \cap B|C) = P(A|C)P(B|C)$ if $A \& B$ indep.

$$= \prod_{x=1}^{n} p(X_i = x_i|\theta)$$

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution

2) Write out the probability of observing our data from this distribution:

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} p(X_i = x_i|\theta)$$

data

We call this the likelihood

Maximum likelihood estimation

We want to estimate the probability distribution that these samples came from.

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution
3) Find the parameter that maximizes the likelihood $\propto P(D|\theta)$

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

value of θ that maximizes $P(D|\theta)$

MLE Properties

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- MLEs are easy to estimate
- MLEs are efficient to estimate
- MLEs have great asymptotic properties

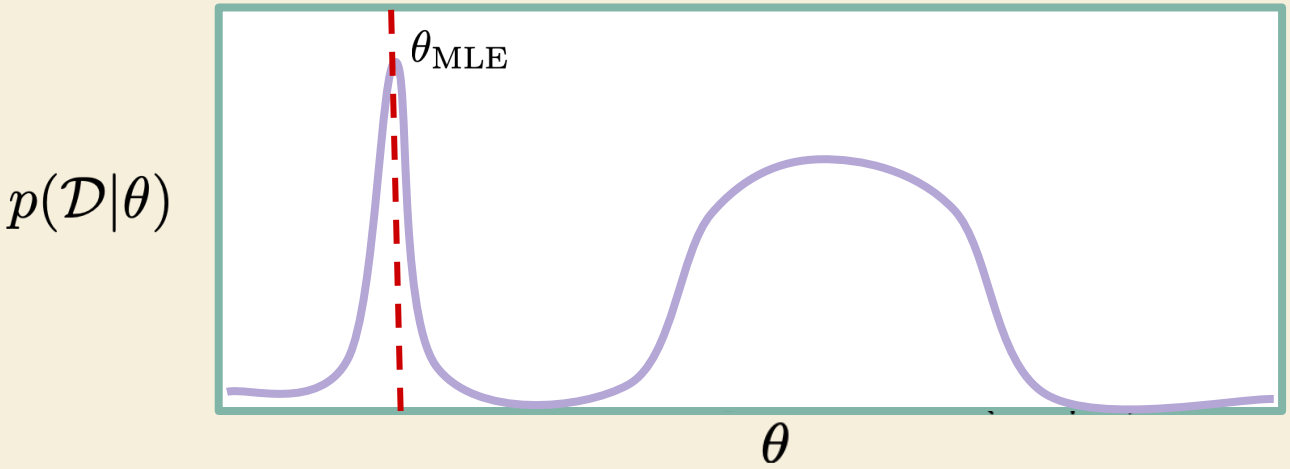
MLE Properties

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- MLEs are not guaranteed to exist
- MLEs are not guaranteed to be unique
- MLEs don't give uncertainty
- MLEs are prone to overfitting

MLE Properties

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$



- *MLEs don't give uncertainty*
- *MLEs are prone to overfitting*

MLE for Bernoulli distribution

Recall the form of a Bernoulli random variable with parameter theta:

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1 - \theta) & \text{if } x = 0 \end{cases}$$

MLE for Bernoulli distribution

Make assumption about form: data dist ~ Bernoulli

Recall the form of a Bernoulli random variable with parameter theta:

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1 - \theta) & \text{if } x = 0 \end{cases}$$

This is the form of our model for a two outcome random variable (e.g., a coin)

MLE for Bernoulli distribution

1) Make an assumption about the form of the distribution - bernoulli

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1 - \theta) & \text{if } x = 0 \end{cases}$$

This is the form of our model for a two outcome random variable (e.g., a coin)

MLE for Bernoulli distribution

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1 - \theta) & \text{if } x = 0 \end{cases}$$

MLE for Bernoulli distribution

Write out the probability of observing our data from this distribution

$$p_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ (1-\theta) & \text{if } x = 0 \end{cases}$$

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

let $x_1 = 1$, $x_2 = 0$, $x_3 = 1$: $P(x_1 = 1|\theta) = \theta$
$P(x_2 = 0|\theta) = 1-\theta$
$P(x_3 = 1|\theta) = \theta$

$P(x_1 = x_1, x_2 = x_2, \ldots)$
$= \prod P(x_i = x_i|\theta)$ as x_i is iid

MLE for Bernoulli distribution

Write out the probability of observing our data from this distribution

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

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} p(X_i = x_i|\theta)$$

$p(X_1, X_2, X_3 \ldots) = p(x_1)\, p(x_2)\, p(x_3) \ldots$

MLE for Bernoulli distribution

Write out the probability of observing our data from this distribution

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

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} p(X_i = x_i|\theta)$$

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

MLE for Bernoulli distribution

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

likelihood

MLE for Bernoulli distribution

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution
3) ???

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

MLE for Bernoulli distribution

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution
3) Find the parameter that maximizes the likelihood

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

$L(\mathcal{D}|\theta) = f(x_1, x_2 \cdots|\theta) = \prod_{i=1} f(x_i|\theta) = \prod_{i=1} \theta^{x_i}(1-\theta)^{1-x_i} = (\theta^{x_1}(1-\theta)^{1-x_1})(\theta^{x_2}(1-\theta)^{1-x_2}) \cdots$

$\log(L(\mathcal{D}|\theta)) = \log[(\theta^{x_1}(1-\theta)^{1-x_1})(\theta^{x_2}(1-\theta)^{1-x_2}) \cdots = \log(\theta^{x_1}(1-\theta)^{1-x_1}) + \log(\theta^{x_2}(1-\theta)^{1-x_2}) + \cdots$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\arg\max_{\theta} \; p(\mathcal{D}|\theta) = \arg\max_{\theta} \; \log\big(p(\mathcal{D}|\theta)\big)$$

Common theme we will see when manipulating likelihoods is that both numerically and practically it can be easier to work with the log likelihood

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$p(X_i = x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} \xrightarrow{\log} x_i \log(\theta) + (1-x_i)\log(1-\theta)$

$$\log\big(p(X_i = x_i|\theta)\big) = \log\big(\theta^{x_i}(1-\theta)^{1-x_i}\big)$$

$$= \log\big(\theta^{x_i}\big) + \log\big((1-\theta)^{1-x_i}\big)$$

$$= x_i \log\big(\theta\big) + (1-x_i)\log\big((1-\theta)\big)$$

$$p(\mathcal{D}|\theta) = \prod_{i=0}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\log\big(p(X_i = x_i|\theta)\big) = \log\big(\theta^{x_i}(1-\theta)^{1-x_i}\big)$$

$$= \log\big(\theta^{x_i}\big) + \log\big((1-\theta)^{1-x_i}\big)$$

$$= x_i \log\big(\theta\big) + (1-x_i)\log\big((1-\theta)\big)$$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\log\big(p(X_i = x_i|\theta)\big) = x_i \log\big(\theta\big) + (1 - x_i)\log\big((1 - \theta)\big)$$

$$\log\big(p(\mathcal{D}|\theta)\big) = \sum_{i=0}^{N} x_i \log\big(\theta\big) + (1 - x_i)\log\big((1 - \theta)\big)$$

$P(D|\theta) = \prod \theta^{x_i}(1-\theta)^{1-x_i} = (\theta^{x_1}(1-\theta)^{1-x_1})(\theta^{x_2}(1-\theta)^{1-x_2})\cdots$

$\log(P(D|\theta)) = \log(\sim) + \log(\sim) + \cdots = x_1 \log\theta + (1-x_1)\log(1-\theta) + \cdots$

$= \sum x_i \log\theta + (1-x_i)\log(1-\theta)$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\log\big(p(\mathcal{D}|\theta)\big) = \sum_{i=0}^{N} x_i \log\big(\theta\big) + (1 - x_i)\log\big((1-\theta)\big)$$

$$\frac{d}{d\theta}\log\big(p(\mathcal{D}|\theta)\big) = \sum_{i=1}^{N}\left(\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}\right)$$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\log\big(p(\mathcal{D}|\theta)\big) = \sum_{i=0}^{N} x_i \log\big(\theta\big) + (1 - x_i)\log\big((1 - \theta)\big)$$

$$\frac{d}{d\theta} \log\left(p(\mathcal{D}|\theta)\right) = \sum_{i=1}^{N} \left(\frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right) = 0$$

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\frac{d}{d\theta}\log\left(p(\mathcal{D}|\theta)\right) = \sum_{i=1}^{N}\left(\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}\right) = 0$$

$$= \sum_{i=1}^{N}\left(x_i(1-\theta) - (1-x_i)\theta\right)$$

Multiply through by theta(1-theta)

MLE for Bernoulli distribution

Find the parameter that maximizes the likelihood

$$\sum_{i=1}^{N} (x_i(1-\theta) - (1-x_i)\theta) = 0$$

$$\sum(x_i - x_i\theta - \theta + x_i\theta) = \sum(x_i - \theta) = \sum x_i - \sum\theta = \sum x_i - N\theta = 0$$

$$\sum x_i = N\theta$$

$$\sum_{i=1}^{N} x_i = N\theta \implies \theta = \boxed{\frac{1}{N}\sum_{i=1}^{n} x_i}$$

Why MLE for linear regression?

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

Until now, our model has been a least squares model where we take the prediction to be the output of a model without any noise. However, it can be very useful to incorporate knowledge about noisy measurements. So it is very common to introduce a noise term:

Why MLE for linear regression?

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

Until now, our model has been a least squares model where we take the prediction to be the output of a model without any noise. However, it can be very useful to incorporate knowledge about noisy measurements. So it is very common to introduce a noise term:

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta + \epsilon$$

Why MLE for linear regression?

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

$$\hat{y}^{(i)} \sim \mathcal{N}(\mathbf{x}^{(i)\top}\theta, \sigma^2\mathbf{I})$$

Given an additive isotropic Gaussian noise model, we can simply recenter our Gaussian at the prediction and now we have a nice form for our probabilistic model

Why MLE for linear regression?

1) Make an assumption about the form of the distribution
2) ???

Assume data come from Gaussian dist (when?)

$$\hat{y}^{(i)} \sim \mathcal{N}(\mathbf{x}^{(i)\top}\theta, \sigma^2\mathbf{I})$$

Why MLE for linear regression?

1) Make an assumption about the form of the distribution
2) Write out the probability of observing our data from this distribution

$$\hat{y}^{(i)} \sim \mathcal{N}(\mathbf{x}^{(i)\top}\theta, \sigma^2\mathbf{I})$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2}{2\sigma^2}\right)$$

Assuming co-domain is just the reals

Computing the MLE in linear regression

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2\right)$$

Computing the MLE in linear regression

$$p(\mathcal{D}|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2\right)$$

$$\log\left(p(\mathcal{D}|\theta)\right) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2$$

Common theme we will see when manipulating likelihoods is that both numerically and practically it can be easier to work with the log likelihood

Computing the MLE in linear regression

$$p(\mathcal{D}|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2\right)$$

$$\log\left(p(\mathcal{D}|\theta)\right) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2$$

Notice! We have recovered something very similar to our MSE loss (MSE is boxed in red). But we have not explicitly stated this loss. That is because the MSE is motivated by this probabilistic model!

Computing the MLE in linear regression

$$p(\mathcal{D}|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2\right)$$

$$\log\left(p(\mathcal{D}|\theta)\right) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2$$

Common theme #2: Maximizing the log likelihood is the same as minimizing the negative log likelihood. Since we are used to minimizing losses in lin. regression, lets do that!

Computing the MLE in linear regression

$$\log\left(p(\mathcal{D}|\theta)\right) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta)^2$$

$$-\log\left(p(\mathcal{D}|\theta)\right) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y})$$

Common theme #2: Maximizing the log likelihood is the same as minimizing the negative log likelihood. Since we are used to minimizing losses in lin. regression, lets do that!
(And also re-write in terms of matrices)

Computing the MLE in linear regression

$$-\log\left(p(\mathcal{D}|\theta)\right) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y})$$

The green term we know yields the OLS estimator we have derived in previous lectures and then it seems the MLE is just a scalar multiple plus a scalar. So the MLE and the OLS have the same value!

Computing the MLE in linear regression

$$-\log\left(p(\mathcal{D}|\theta)\right) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y})$$

Mean Squared error: $\frac{1}{2N}(X\theta-y)^T(X\theta-y)$

The green term we know yields the OLS estimator we have derived in previous lectures and then it seems the MLE is just a scalar multiple plus a scalar. So the MLE and the OLS have the same value!

$(\theta^T X^T - y^T)(X\theta - y) =$
$\theta^T X^T X\theta - 2y^T X\theta - y^T y$

$\frac{\partial}{\partial\theta}(*) : 2\theta^T X^T X - 2y^T X = 0$
$(\theta^T X^T X)^T = (y^T X)^T$
$X^T X\theta = X^T y$
$\theta = (X^T X)^{-1} X^T y = \theta_{MLE}$

What should we pick for sigma?

$\theta_{MLE} = (X^T X)^{-1} X^T y$

Computing the MLE in linear regression

$$-\log\left(p(\mathcal{D}|\theta)\right) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\theta - \mathbf{y})^\top(\mathbf{X}\theta - \mathbf{y})$$

What should we pick for sigma? (Left as additional exercise)

$$\sigma^2_{\text{MLE}} = \frac{1}{N}(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta_{\text{MLE}})^\top(\hat{y}^{(i)} - \mathbf{x}^{(i)\top}\theta_{\text{MLE}})$$

Break!

A quick look at classification

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\} \qquad \mathbf{x} \in \mathbb{R}^n \qquad y \in \{0, 1\}$$

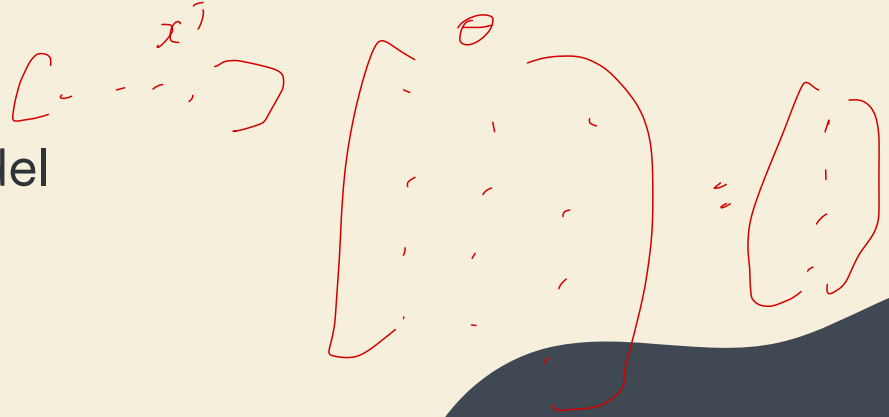
If we want to use our linear model, what loss should we pick?

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\} \qquad \mathbf{x} \in \mathbb{R}^n \qquad y \in \{0, 1\}$$

$$\mathbf{x}^\top \theta$$

We know the form of our model

And our output looks a lot like the coin flip we just saw!



If we want to use our linear model, what loss should we pick?

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\} \qquad \mathbf{x} \in \mathbb{R}^n \qquad y \in \{0, 1\}$$

And our output looks a lot like the coin flip we just saw!

$$\mathbf{x}^\top \theta$$

We know the form of our model

If we want to use our linear model, what loss should we pick?

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\} \qquad \mathbf{x} \in \mathbb{R}^n \qquad y \in \{0, 1\}$$

$$\text{sigmoid}(\mathbf{x}^\top \theta)$$

sigmoid: $\frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

We can modify the output of our model to be scaled between zero and one

$0 \leq \text{sig}(x^T\theta) \leq 1$

Logistic regression

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\} \qquad \mathbf{x} \in \mathbb{R}^n \qquad y \in \{0, 1\}$$

$$p(y|\mathbf{x}, \theta) = \begin{cases} \text{sigmoid}(\mathbf{x}^\top \theta) & \text{we have } y = 1 \\ (1 - \text{sigmoid}(\mathbf{x}^\top \theta)) & \text{we have } y = 0 \end{cases}$$

We take the output of the scaled model to be the parameter of our Bernoulli likelihood from before

Logistic regression

$$p(y|\mathbf{x},\theta) = \begin{cases} \text{sigmoid}(\mathbf{x}^\top\theta) & \text{we have } y=1 \\ (1-\text{sigmoid}(\mathbf{x}^\top\theta)) & \text{we have } y=0 \end{cases}$$

We know from the start of lecture that the log of the Bernoulli likelihood gives us:

$$\log\left(p(y|\mathbf{x},\theta)\right) = y\log(\text{sigm}(\mathbf{x}^\top\theta)) + (1-y)(1-\log(\text{sigm}(\mathbf{x}^\top\theta)))$$

This is the cross-entropy loss that we use today in almost all classification tasks! (Again, we typically use the negative log likelihood)

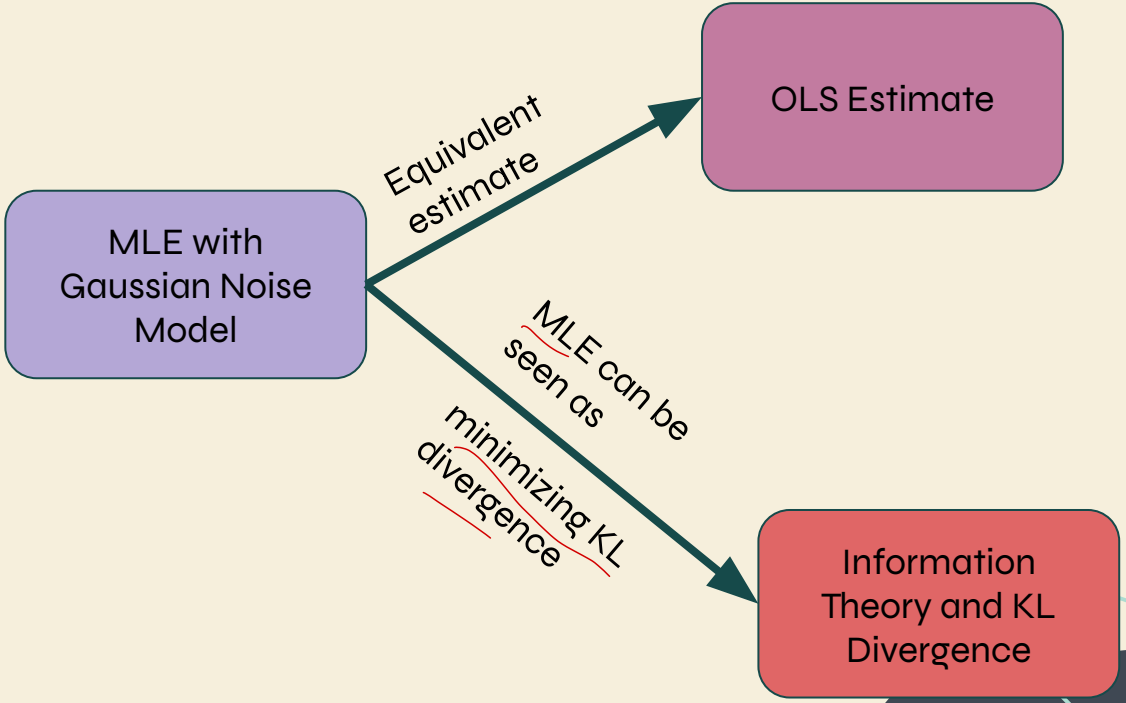
Logistic regression

$$p(y|\mathbf{x},\theta) = \begin{cases} \text{sigmoid}(\mathbf{x}^\top\theta) & \text{we have } y=1 \\ (1-\text{sigmoid}(\mathbf{x}^\top\theta)) & \text{we have } y=0 \end{cases}$$

$$\log\left(p(y|\mathbf{x},\theta)\right) = y\log(\text{sigm}(\mathbf{x}^\top\theta)) + (1-y)(1-\log(\text{sigm}(\mathbf{x}^\top\theta)))$$

Take away: forming a probabilistic model allows us to easily identify good objectives for our learning models

Exploring Further MLE connections



Information Theory

Let X be a random variable with p.m.f. p. The entropy of the random variable can be defined:

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Focusing again on Bernoulli random variables:

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

Information Theory

Let X be a random variable with p.m.f. p. The entropy of the random variable can be defined:

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Focusing again on Bernoulli random variables:

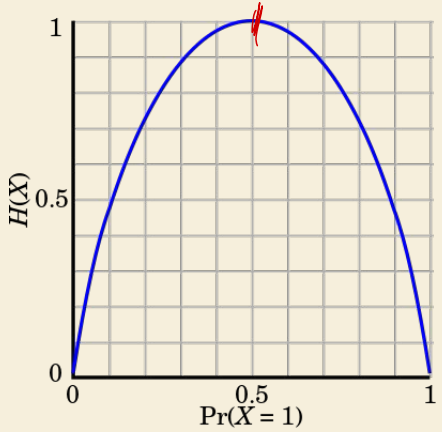
$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Information Theory

Focusing again on Bernoulli random variables:

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

$$H(X) = -\theta \log_2(\theta) - (1-\theta)\log_2(1-\theta)$$



Information Theory: KL Divergence

Taking things one step further, we have the KL divergence is an information theoretic "distance" between distributions (n.b. it is not a distance, it is a divergence)

$$\mathrm{KL}(p||q) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

Information Theory: KL Divergence

Taking things one step further, we have the KL divergence:

$$\mathrm{KL}(p||q) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

$$= \sum_{x \in X} p(x) \log(p(x)) - \sum_{x \in X} p(x) \log(q(x))$$

$$\boxed{= -H(p) + H(p, q)}$$

This is called the cross entropy (you can check that this matches the general form we have before!)

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log(q(x_i|\theta))$$

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log(q(x_i|\theta))$$

$$= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i|\theta)) - \frac{1}{N} \sum_{i=1}^{N} p(x_i)$$

We can change to an average because it is simply multiplying by a constant and we subtracted a value that does not depend on theta at all

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log(q(x_i|\theta))$$

$$= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i|\theta)) - \frac{1}{N} \sum_{i=1}^{N} p(x_i)$$

$$= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{p(x_i)}{q(x_i|\theta)}\right)$$

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log(q(x_i|\theta))$$

$$= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i|\theta)) - \frac{1}{N} \sum_{i=1}^{N} p(x_i)$$

$$= \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{p(x_i)}{q(x_i|\theta)}\right)$$

$$\rightarrow \arg\min_{\theta} \text{KL}(p \,||\, q_\theta)$$

Linking KL to MLE

$$\theta_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} q(x_i|\theta) \quad \text{(Maximum Likelihood)}$$

$$\rightarrow \arg\min_{\theta} \text{KL}(p \,||\, q_\theta)$$

The link between ML and information theory will not be examined, but I think serves as an important motivation for our emphasis on probabilistic methods: it unlocks tools from a wide range of math subfields that can be used to great effect!

Motivating maximum a posteriori

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- MLEs are not guaranteed to exist
- MLEs are not guaranteed to be unique
- MLEs don't give uncertainty
- *MLEs are prone to overfitting*

Maximum a posteriori

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- MLEs are not guaranteed to exist
- MLEs are not guaranteed to be unique
- MLEs don't give uncertainty
- *MLEs are prone to overfitting*

Maximum a posteriori

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

- Start by assuming a the data comes from a joint distribution: $p(\theta, \mathcal{D})$
- The critical thing to notice here is this treats our parameters as a random variable

Maximum a posteriori

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

- Start by assuming a the data comes from a joint distribution: $p(\theta, \mathcal{D})$
- The critical thing to notice here is this treats our parameters as a random variable
- If it is a random variable, then we should reason about its distribution: $p(\theta)$

Maximum a posteriori

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

- Start by assuming a the data comes from a joint distribution: $p(\theta, \mathcal{D})$
- Using the product rule from probability we know:

$$p(\theta, \mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)$$

MLE only uses likelihood
MAP uses likelihood & prior

Maximum a posteriori

$$\mathcal{D} := \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$$

$$\hat{y}^{(i)} = \mathbf{x}^{(i)\top}\theta$$

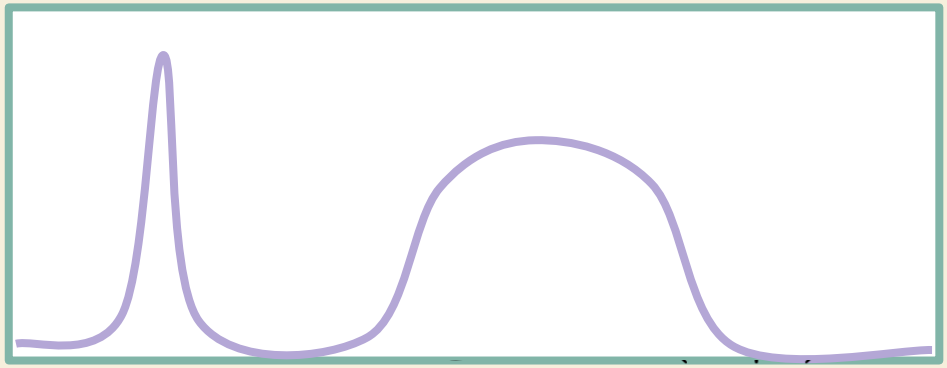
$p(x, y) = p(x|y)p(y)$

$$p(\theta, \mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)$$

joint distribution

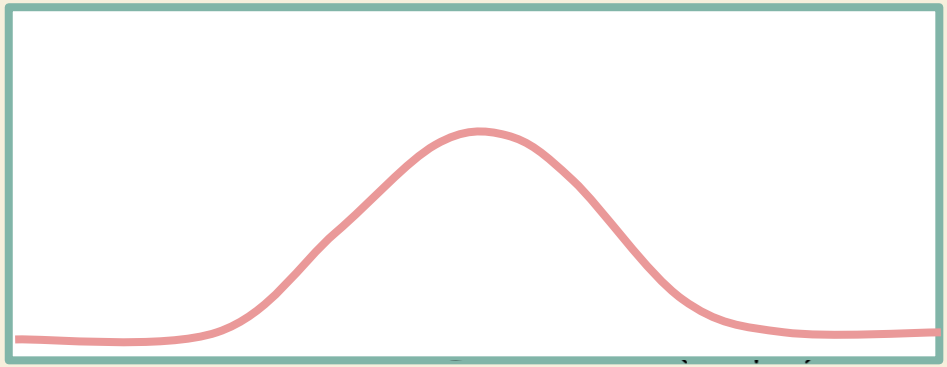
The prior distribution allows us to express some knowledge we have about our model. This can in turn help us avoid overfitting.

$p(\mathcal{D}|\theta)$



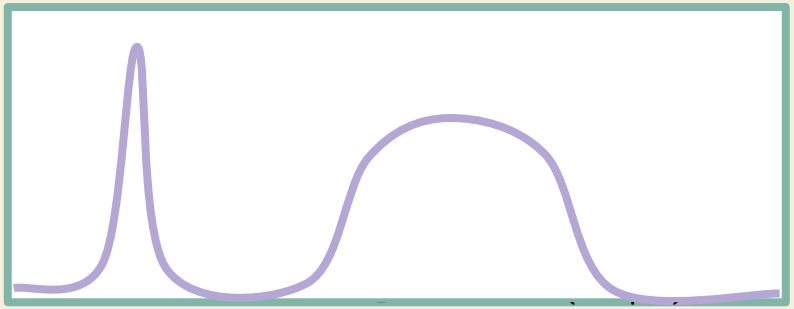
θ

$p(\theta)$



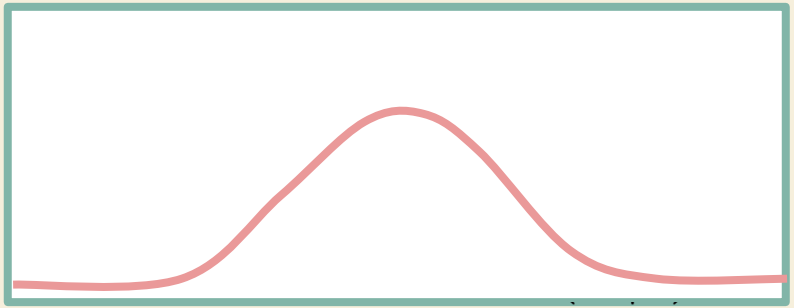
θ

$p(\mathcal{D}|\theta)$



θ

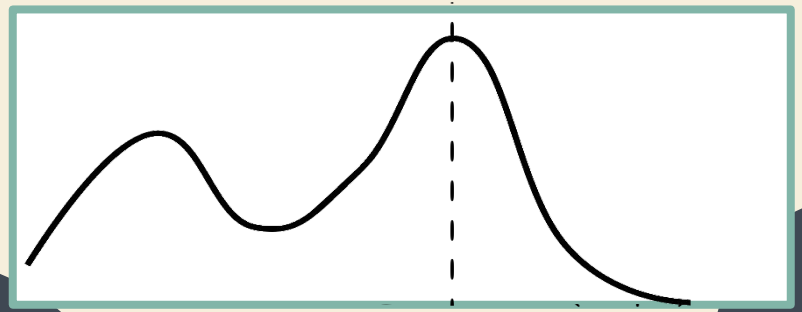
$p(\theta)$



θ

Notice this is not exactly the joint distribution we had before. We call this the posterior distribution

$p(\theta|\mathcal{D})$

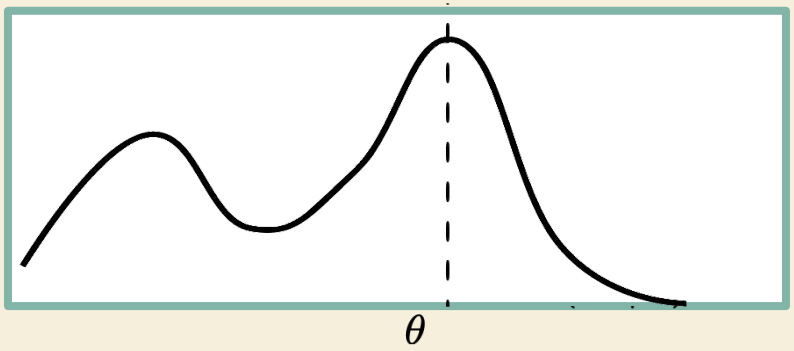


θ

Maximum a posteriori

Notice this is not exactly the joint distribution we had before. We call this the posterior distribution

$p(\theta|\mathcal{D})$



We will deep dive into the posterior distribution from before, but let us use the same principles from MLE to derive the maximum of the posterior distribution which we call the MAP

Maximum a posteriori

$$\theta_{\text{MAP}} = \arg\max_{\theta} \; p(\theta|\mathcal{D})$$

$$= \arg\max_{\theta} \; p(\mathcal{D}|\theta)p(\theta)$$

Maximum a posteriori

$$\theta_{\text{MAP}} = \arg\max_{\theta} \; p(\theta|\mathcal{D})$$

$$= \arg\max_{\theta} \; p(\mathcal{D}|\theta)p(\theta)$$

$$p(\mathcal{D}|\theta)p(\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2} \right) \mathcal{N}(\theta; 0, \tau^2\mathbf{I})$$

mean = 0

P(D|θ) from MLE

P(θ)

assume 0 mean so no bias

First we expand the likelihood term (this is the same as in our MLE derivation)

Maximum a posteriori

$$p(\mathcal{D}|\theta)p(\theta)$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2}\right)\mathcal{N}(\theta;0,\tau^2\mathbf{I})$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2}\right)\frac{1}{(2\pi\tau^2)^{d/2}}\exp\left(-\frac{\theta^\top\theta}{2\tau^2}\right)$$

Then we expand our prior distribution to get our the full equation that dictates

Maximum a posteriori

$$\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2}\right)\frac{1}{(2\pi\tau^2)^{d/2}}\exp\left(-\frac{\theta^\top\theta}{2\tau^2}\right)$$

Now to find the argmax parameter for this model we need to follow our steps from our MLE exposition:

1. Go from this likelihood to the negative log likelihood (NLL)
2. Set equal to zero and solve for theta

Maximum a posteriori

$$\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{X}\theta)^\top(\mathbf{y}-\mathbf{X}\theta)}{2\sigma^2}\right)\frac{1}{(2\pi\tau^2)^{d/2}}\exp\left(-\frac{\theta^\top\theta}{2\tau^2}\right)$$

Now to find the argmax parameter for this model we need to follow our steps from our MLE exposition:

1. Go from this likelihood to the negative log likelihood (NLL)
2. Set equal to zero and solve for theta

$$\theta^{\text{MAP}} = \left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Maximum a posteriori

$$\theta_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\theta_{\text{MAP}} = \left(\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Let's take a second to reason about how this is different from our MLE estimate!

As we increase tau to infinity (interpret as making the prior infinitely wide and thus uninformative) we get back the MLE estimate

As we decrease tau to an infinitesimally small value (indicating a strong prior distribution), the effect is that the inverse converges to zero!

Maximum a posteriori

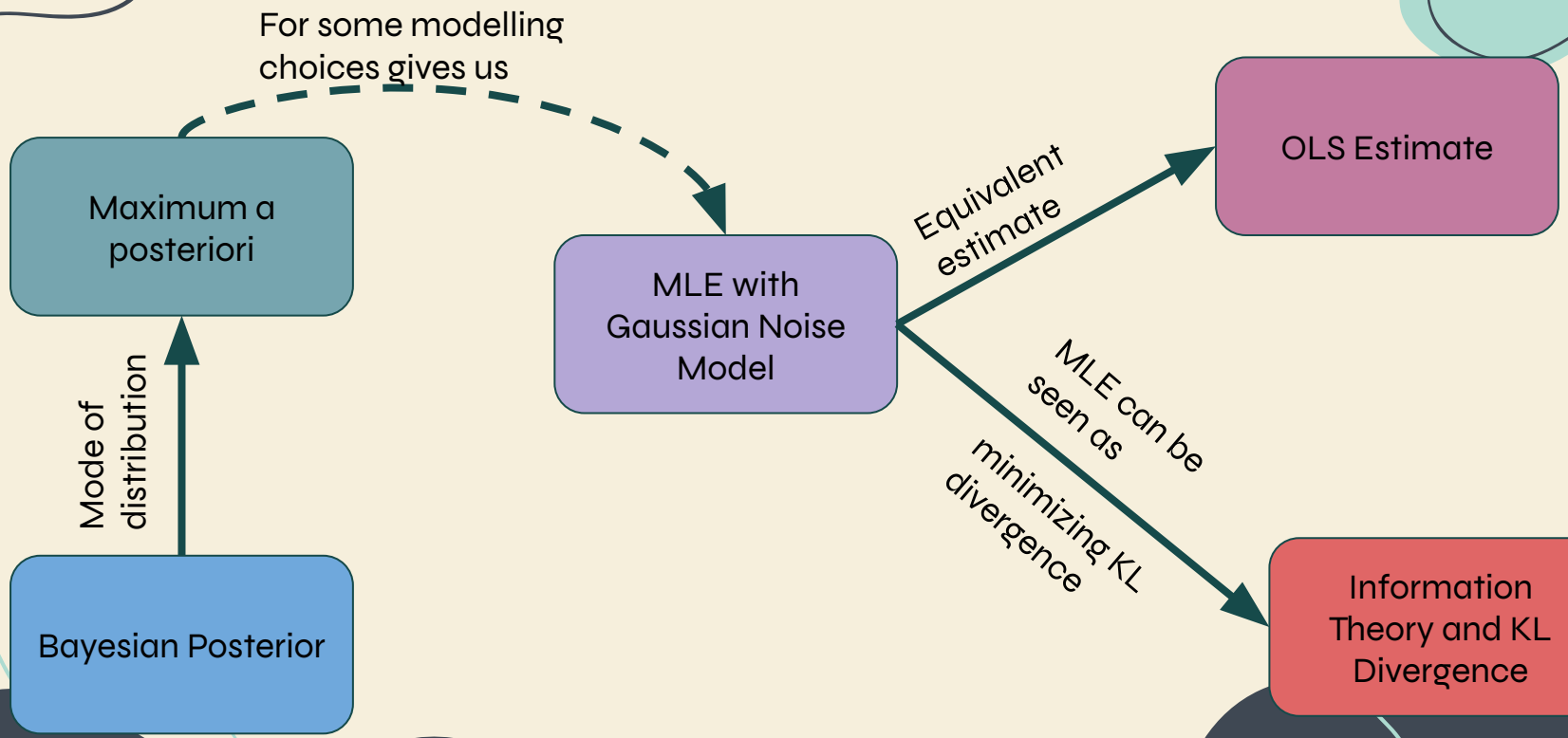
$$\theta_{\text{MAP}} = \left(\mathbf{X}^\top \mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

So how exactly should we pick tau? This depends on the modelling circumstance and can be seen as either a drawback or a strength of the Bayesian modelling framework.

Another disadvantage of the MAP (that is shared with the MLE) is that it is only a point estimate, giving us no way to quantify uncertainty.

Before providing the full Bayesian inference framework, we will do a review of key multivariate probability skills.

Exploring Further MLE connections



Next lecture: probability theory review!