

ELEN 4720: Machine Learning for Signals, Information and Data

Lecture 15

Prof. John Paisley

Department of Electrical Engineering

Columbia University

MAXIMUM LIKELIHOOD

APPROACHES TO DATA MODELING

Our approaches to modeling data thus far have been either probabilistic or non-probabilistic in motivation.

- ▶ Probabilistic models: Probability distributions defined on data, e.g.,
 1. Bayes classifiers
 2. Logistic regression
 3. Least squares and ridge regression (using ML and MAP interpretation)
 4. Bayesian linear regression
- ▶ Non-probabilistic models: No probability distributions involved, e.g.,
 1. Perceptron
 2. Support vector machine
 3. Decision trees
 4. K-means

In *every* case, we have some objective function we are trying to optimize (greedily vs non-greedily, locally vs globally).

MAXIMUM LIKELIHOOD

As we've seen, one *probabilistic* objective function is maximum likelihood.

Setup: In the most basic scenario, we start with

1. some set of model parameters θ
2. a set of data $\{x_1, \dots, x_n\}$
3. a probability distribution $p(x|\theta)$
4. an i.i.d. assumption, $x_i \stackrel{iid}{\sim} p(x|\theta)$

Maximum likelihood seeks the θ that maximizes the likelihood

$$\theta_{\text{ML}} = \arg \max_{\theta} p(x_1, \dots, x_n | \theta) \stackrel{(a)}{=} \arg \max_{\theta} \prod_{i=1}^n p(x_i | \theta) \stackrel{(b)}{=} \arg \max_{\theta} \sum_{i=1}^n \ln p(x_i | \theta)$$

(a) follows from i.i.d. assumption.

(b) follows since $f(y) > f(x) \Rightarrow \ln f(y) > \ln f(x)$.

MAXIMUM LIKELIHOOD

We've discussed maximum likelihood for a few models, e.g., least squares linear regression and the Bayes classifier.

Both of these models were “nice” because we could find their respective θ_{ML} analytically by writing an equation and plugging in data to solve.

Gaussian with unknown mean and covariance

In the first lecture, we saw if $x_i \stackrel{iid}{\sim} N(\mu, \Sigma)$, where $\theta = \{\mu, \Sigma\}$, then

$$\nabla_{\theta} \ln \prod_{i=1}^n p(x_i|\theta) = 0$$

gives the following maximum likelihood values for μ and Σ :

$$\mu_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \Sigma_{\text{ML}} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\text{ML}})(x_i - \mu_{\text{ML}})^T$$

COORDINATE ASCENT AND MAXIMUM LIKELIHOOD

In more complicated models, we might split the parameters into groups θ_1, θ_2 and try to maximize the likelihood over both of these,

$$\theta_{1,\text{ML}}, \theta_{2,\text{ML}} = \arg \max_{\theta_1, \theta_2} \sum_{i=1}^n \ln p(x_i | \theta_1, \theta_2),$$

Although we can solve one *given* the other, we can't solve it *simultaneously*.

Coordinate ascent (probabilistic version)

We saw how K-means presented a similar situation, and that we could optimize using this method. This technique is generalizable.

Algorithm: For iteration $t = 1, 2, \dots$,

1. Optimize $\theta_1^{(t)} = \arg \max_{\theta_1} \sum_{i=1}^n \ln p(x_i | \theta_1, \theta_2^{(t-1)})$
2. Optimize $\theta_2^{(t)} = \arg \max_{\theta_2} \sum_{i=1}^n \ln p(x_i | \theta_1^{(t)}, \theta_2)$

COORDINATE ASCENT AND MAXIMUM LIKELIHOOD

There is a third (subtly) different situation, where we really want to find

$$\theta_{1,\text{ML}} = \arg \max_{\theta_1} \sum_{i=1}^n \ln p(x_i | \theta_1).$$

Except this function is “tricky” to optimize directly. However, we figure out that we can add a second variable θ_2 to the model such that

$$\sum_{i=1}^n \ln p(x_i, \theta_2 | \theta_1) \quad (\text{Function 2})$$

is easier to work with. We’ll make this clearer later.

- ▶ Notice in this second case that θ_2 is on the *left* side of the conditioning bar. This implies a prior on θ_2 , (whatever “ θ_2 ” turns out to be).
- ▶ We will next discuss a fundamental technique called the EM algorithm for finding $\theta_{1,\text{ML}}$ by using Function 2 instead.

EXPECTATION-MAXIMIZATION ALGORITHM

A MOTIVATING EXAMPLE

Let $x_i \in \mathbb{R}^d$, be a vector with *missing data*. Split this vector into two parts:

1. x_i^o – observed portion (the sub-vector of x_i that is measured)
2. x_i^m – missing portion (the sub-vector of x_i that is still unknown)
3. The missing dimensions can be different for different x_i .

We assume that $x_i \stackrel{iid}{\sim} N(\mu, \Sigma)$, and want to solve

$$\mu_{\text{ML}}, \Sigma_{\text{ML}} = \arg \max_{\mu, \Sigma} \sum_{i=1}^n \ln p(x_i^o | \mu, \Sigma).$$

This is tricky. However, if we knew x_i^m (and therefore x_i), then

$$\mu_{\text{ML}}, \Sigma_{\text{ML}} = \arg \max_{\mu, \Sigma} \sum_{i=1}^n \underbrace{\ln p(x_i^o, x_i^m | \mu, \Sigma)}_{= p(x_i | \mu, \Sigma)}$$

is very easy to optimize (we just did it on a previous slide).

CONNECTING TO A MORE GENERAL SETUP

We will discuss a method for optimizing $\sum_{i=1}^n \ln p(x_i^o | \mu, \Sigma)$ and imputing its missing values $\{x_1^m, \dots, x_n^m\}$. This is a very general technique.

General setup

Imagine we have two parameter sets θ_1, θ_2 , where

$$p(x|\theta_1) = \int p(x, \theta_2|\theta_1) d\theta_2 \quad (\text{marginal distribution})$$

Example: For the previous example we can show that

$$p(x_i^o | \mu, \Sigma) = \int p(x_i^o, x_i^m | \mu, \Sigma) dx_i^m = N(\mu_i^o, \Sigma_i^o),$$

where μ_i^o and Σ_i^o are the sub-vector/sub-matrix of μ and Σ defined by x_i^o .

THE EM OBJECTIVE FUNCTION

We need to define a general *objective function* that gives us what we want:

1. It lets us optimize the marginal $p(x|\theta_1)$ over θ_1 ,
2. It uses $p(x, \theta_2|\theta_1)$ in doing so purely for computational convenience.

The EM objective function

Before picking it apart, we claim that this objective function is

$$\ln p(x|\theta_1) = \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2 + \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2$$

Some immediate comments:

- ▶ $q(\theta_2)$ is *any* probability distribution (assumed continuous for now)
- ▶ We assume we know $p(\theta_2|x, \theta_1)$. That is, given the data x and fixed values for θ_1 , we can solve the conditional posterior distribution of θ_2 .

DERIVING THE EM OBJECTIVE FUNCTION

Let's show that this equality is actually true

$$\begin{aligned}\ln p(x|\theta_1) &= \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2 + \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2 \\ &= \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)q(\theta_2)}{p(\theta_2|x, \theta_1)q(\theta_2)} d\theta_2\end{aligned}$$

Remember some rules of probability:

$$p(a, b|c) = p(a|b, c)p(b|c) \quad \Rightarrow \quad p(b|c) = \frac{p(a, b|c)}{p(a|b, c)}.$$

Letting $a = \theta_2$, $b = x$ and $c = \theta_1$, we conclude

$$\begin{aligned}\ln p(x|\theta_1) &= \int q(\theta_2) \ln p(x|\theta_1) d\theta_2 \\ &= \ln p(x|\theta_1)\end{aligned}$$

THE EM OBJECTIVE FUNCTION

The EM objective function splits our desired objective into two terms:

$$\ln p(x|\theta_1) = \underbrace{\int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2}_{\text{A function only of } \theta_1, \text{ we'll call it } \mathcal{L}} + \underbrace{\int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2}_{\text{Kullback-Leibler divergence}}$$

Some more observations about the right hand side:

1. The **KL divergence** is always ≥ 0 and only $= 0$ when $q = p$.
2. We are assuming that the integral in $\mathcal{L}$ can be calculated, leaving a function only of θ_1 (for a particular setting of the distribution q).

BIGGER PICTURE

Q: What does it mean to iteratively optimize $\ln p(x|\theta_1)$ w.r.t. θ_1 ?

A: One way to think about it is that we want a method for generating:

1. A sequence of values for θ_1 such that $\ln p(x|\theta_1^{(t)}) \geq \ln p(x|\theta_1^{(t-1)})$.
2. We want $\theta_1^{(t)}$ to converge to a local maximum of $\ln p(x|\theta_1)$.

It doesn't matter how we generate the sequence $\theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(3)}, \dots$

We will show how EM generates #1 and just mention that EM satisfies #2.

THE EM ALGORITHM

The EM objective function

$$\ln p(x|\theta_1) = \underbrace{\int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2}_{\text{define this to be } \mathcal{L}(x, \theta_1)} + \underbrace{\int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2}_{\text{Kullback-Leibler divergence}}$$

Definition: The EM algorithm

Given the value $\theta_1^{(t)}$, **find** the value $\theta_1^{(t+1)}$ as follows:

E-step: Set $q_t(\theta_2) = p(\theta_2|x, \theta_1^{(t)})$ and calculate

$$\mathcal{L}_{q_t}(x, \theta_1) = \int q_t(\theta_2) \ln p(x, \theta_2|\theta_1) d\theta_2 - \underbrace{\int q_t(\theta_2) \ln q_t(\theta_2) d\theta_2}_{\text{can ignore this term}}.$$

M-step: Set $\theta_1^{(t+1)} = \arg \max_{\theta_1} \mathcal{L}_{q_t}(x, \theta_1)$.

PROOF OF MONOTONIC IMPROVEMENT (OPTIONAL)

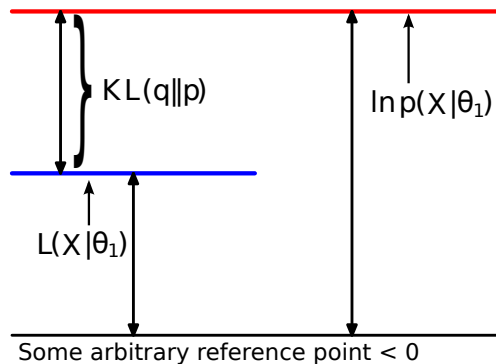
Once we're comfortable with the moving parts, the proof that the sequence $\theta_1^{(t)}$ monotonically improves $\ln p(x|\theta_1)$ just requires *analysis*:

$$\begin{aligned}\ln p(x|\theta_1^{(t)}) &= \mathcal{L}(x, \theta_1^{(t)}) + \underbrace{KL\left(q(\theta_2) \parallel p(\theta_2|x_1, \theta_1^{(t)})\right)}_{= 0 \text{ by setting } q = p} \\ &= \mathcal{L}_{q_t}(x, \theta_1^{(t)}) \quad \leftarrow \text{E-step} \\ &\leq \mathcal{L}_{q_t}(x, \theta_1^{(t+1)}) \quad \leftarrow \text{M-step} \\ &\leq \mathcal{L}_{q_t}(x, \theta_1^{(t+1)}) + \underbrace{KL\left(q_t(\theta_2) \parallel p(\theta_2|x_1, \theta_1^{(t+1)})\right)}_{> 0 \text{ because } q \neq p}\end{aligned}$$

$$\begin{aligned}\text{generalize to any } q \rightarrow &= \mathcal{L}(x, \theta_1^{(t+1)}) + KL\left(q(\theta_2) \parallel p(\theta_2|x_1, \theta_1^{(t+1)})\right) \\ &= \ln p(x|\theta_1^{(t+1)})\end{aligned}$$

ONE ITERATION OF EM (OPTIONAL)

Start: Current setting of θ_1 and $q(\theta_2)$



For reference:

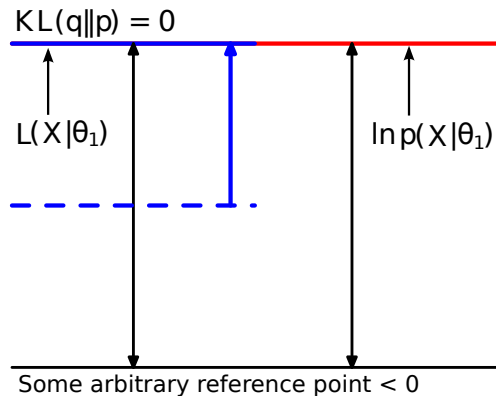
$$\ln p(x|\theta_1) = \mathcal{L} + KL$$

$$\mathcal{L} = \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2$$

$$KL = \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2$$

ONE ITERATION OF EM (OPTIONAL)

E-step: Set $q(\theta_2) = p(\theta_2|x, \theta_1)$ and update $\mathcal{L}$.



For reference:

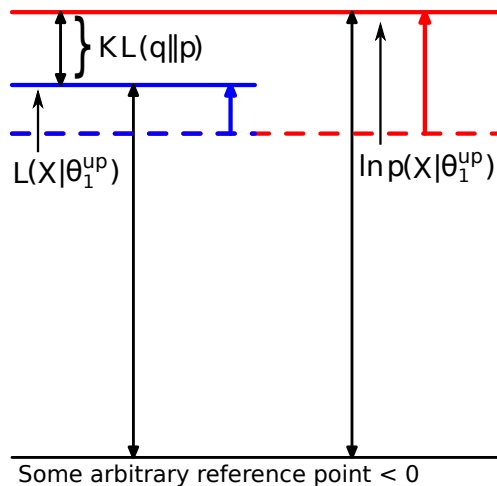
$$\ln p(x|\theta_1) = \mathcal{L} + KL$$

$$\mathcal{L} = \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2$$

$$KL = \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2$$

ONE ITERATION OF EM (OPTIONAL)

M-step: Maximize $\mathcal{L}$ wrt θ_1 . Now $q \neq p$.



For reference:

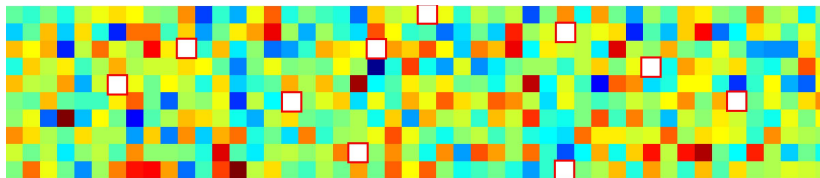
$$\ln p(x|\theta_1) = \mathcal{L} + KL$$

$$\mathcal{L} = \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2$$

$$KL = \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2$$

EM FOR MISSING DATA

THE PROBLEM



We have a data matrix with missing entries. We model the columns as

$$x_i \stackrel{iid}{\sim} N(\mu, \Sigma).$$

Our goal could be to

1. Learn μ and Σ using maximum likelihood
2. Fill in the missing values “intelligently” (e.g., using a model)
3. Both

We will see how to achieve both of these goals using the EM algorithm.

EM FOR SINGLE GAUSSIAN MODEL WITH MISSING DATA

The original, generic EM objective is

$$\ln p(x|\theta_1) = \int q(\theta_2) \ln \frac{p(x, \theta_2|\theta_1)}{q(\theta_2)} d\theta_2 + \int q(\theta_2) \ln \frac{q(\theta_2)}{p(\theta_2|x, \theta_1)} d\theta_2$$

The EM objective for this specific problem and notation is

$$\begin{aligned} \sum_{i=1}^n \ln p(x_i^o | \mu, \Sigma) &= \sum_{i=1}^n \int q(x_i^m) \ln \frac{p(x_i^o, x_i^m | \mu, \Sigma)}{q(x_i^m)} dx_i^m + \\ &\quad \sum_{i=1}^n \int q(x_i^m) \ln \frac{q(x_i^m)}{p(x_i^m | x_i^o, \mu, \Sigma)} dx_i^m \end{aligned}$$

We can calculate everything required to do this.

E-STEP (PART ONE)

Set $q(x_i^m) = p(x_i^m | x_i^o, \mu, \Sigma)$ using current μ, Σ

Let x_i^o and x_i^m represent the observed and missing dimensions of x_i . For notational convenience, think

$$x_i = \begin{bmatrix} x_i^o \\ x_i^m \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_i^o \\ \mu_i^m \end{bmatrix}, \begin{bmatrix} \Sigma_i^{oo} & \Sigma_i^{om} \\ \Sigma_i^{mo} & \Sigma_i^{mm} \end{bmatrix} \right)$$

Then we can show that $p(x_i^m | x_i^o, \mu, \Sigma) = N(\hat{\mu}_i, \hat{\Sigma}_i)$, where

$$\hat{\mu}_i = \mu_i^m + \Sigma_i^{mo} (\Sigma_i^{oo})^{-1} (x_i^o - \mu_i^o), \quad \hat{\Sigma}_i = \Sigma_i^{mm} - \Sigma_i^{mo} (\Sigma_i^{oo})^{-1} \Sigma_i^{om}.$$

It doesn't look nice, but these are just functions of sub-vectors of μ and sub-matrices of Σ using the relevant dimensions defined by x_i .

E-STEP (PART TWO)

E-step: $\mathbb{E}_{q(x_i^m)} [\ln p(x_i^o, x_i^m | \mu, \Sigma)]$

For each i we will need to calculate the following term,

$$\begin{aligned}\mathbb{E}_q[(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)] &= \mathbb{E}_q[\text{trace}\{\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T\}] \\ &= \text{trace}\{\Sigma^{-1} \mathbb{E}_q[(x_i - \mu)(x_i - \mu)^T]\}\end{aligned}$$

The expectation is calculated using $q(x_i^m) = p(x_i^m | x_i^o, \mu, \Sigma)$. So only the missing x_i^m portion of x_i will be integrated over.

To this end, recall $q(x_i^m) = N(\hat{\mu}_i, \hat{\Sigma}_i)$. We define

1. $\hat{x}_i$: A vector where we replace the missing values in x_i with $\hat{\mu}_i$.
2. $\hat{V}_i$: A matrix of 0's, plus sub-matrix $\hat{\Sigma}_i$ in the missing dimensions.

M-STEP

M-step: Maximize $\sum_{i=1}^n \mathbb{E}_q[\ln p(x_i^o, x_i^m | \mu, \Sigma)]$

We'll omit the derivation, but the expectation can now be solved and

$$\mu_{\text{up}}, \Sigma_{\text{up}} = \arg \max_{\mu, \Sigma} \sum_{i=1}^n \mathbb{E}_q[\ln p(x_i^o, x_i^m | \mu, \Sigma)]$$

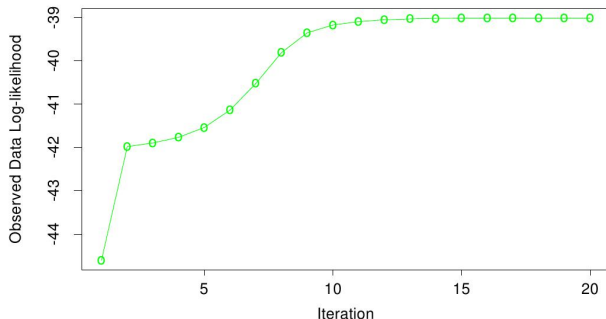
can be found. Recalling the $\hat{}$ notation,

$$\mu_{\text{up}} = \frac{1}{n} \sum_{i=1}^n \hat{x}_i,$$

$$\Sigma_{\text{up}} = \frac{1}{n} \sum_{i=1}^n \{(\hat{x}_i - \mu_{\text{up}})(\hat{x}_i - \mu_{\text{up}})^T + \hat{V}_i\}$$

Then return to the E-step to calculate the new $p(x_i^m | x_i^o, \mu_{\text{up}}, \Sigma_{\text{up}})$.

IMPLEMENTATION DETAILS



We need to initialize μ and Σ , for example, by setting missing values to zero and calculating μ_{ML} and Σ_{ML} . (We can also use random initialization.)

The EM objective function is then calculated after each update to μ and Σ and will look like the figure above. Stop when the change is “small.”

The output is μ_{ML} , Σ_{ML} and $q(x_i^m)$ for all missing entries.