COMPSCI 689
Lecture 4: Linear Regression

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

Maximum Likelihood Estimation

- The maximum likelihood principle asserts that the optimal parameters θ of a probability model $P(\mathbf{Z}|\theta)$ are the parameters that make the observed data the most likely.

- Maximum Likelihood Estimation is a method for selecting the parameters θ of a parametric probability model $P(\mathbf{Z}|\theta)$ by maximizing the (log) likelihood function:

$$\begin{aligned}\theta_* &= \arg\max_{\theta} l(\mathcal{D}, \theta) \\ &= \arg\max_{\theta} \sum_{n=1}^{N} \log P(\mathbf{Z} = \mathbf{z}_n|\theta)\end{aligned}$$

Maximum Conditional Likelihood Estimation

- A slight modification allows us to apply the Maximum Likelihood Principle to conditional probability models of the form $P(\mathbf{Y}|\mathbf{X}=\mathbf{x},\theta)$.

- Maximum Conditional Likelihood Estimation is a method for selecting the parameters θ of a parametric conditional probability model $P(\mathbf{Y}|\mathbf{X}=\mathbf{x},\theta)$ by maximizing the (log) conditional likelihood function:

$$\theta_* = \arg\max_{\theta} l(\mathcal{D},\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log P(\mathbf{Y}=\mathbf{y}_n|\mathbf{X}=\mathbf{x}_n,\theta)$$

- This is the basis for maximum likelihood-based supervised learning methods that attempt to directly approximate $P_*(\mathbf{Y}=\mathbf{y}_n|\mathbf{X}=\mathbf{x}_n)$ using a parametric model $P(\mathbf{Y}|\mathbf{X},\theta)$.

Linear Gaussian Models

- Suppose $y \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^D$.
- Let $\theta = [\mathbf{w}, b, \sigma]$ where $\mathbf{w} \in \mathbb{R}^D$, $\mathbf{b} \in \mathbb{R}$ and $\sigma \in \mathbb{R}^{>0}$.
- A linear Gaussian model has the form:

$$
\begin{aligned}
P(\mathbf{Y} = y | \mathbf{X} = x, \theta) &= \mathcal{N}(y; \mathbf{w}\mathbf{x}^T + b, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - (\mathbf{w}\mathbf{x}^T + b))^2\right)
\end{aligned}
$$

- $\mathbf{w}$ are referred to as the weights, b is the bias, and σ is the standard deviation of the noise.

Absorbing the Bias

- A common trick for simplifying this model is to absorb the bias into the weights and add an extra “1" to the feature vector:

$$\mathbf{w}\mathbf{x}^T + b = [\mathbf{w}, b] \cdot [\mathbf{x}, 1]^T = \tilde{w}\tilde{x}^T$$

- We will assume that the bias has been absorbed to simplify the subsequent derivations. The model we will work with is thus:

$$\begin{aligned} P(\mathbf{Y} = y|\mathbf{X} = x, \theta) &= \mathcal{N}(y; \mathbf{w}\mathbf{x}^T, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mathbf{w}\mathbf{x}^T)^2\right) \end{aligned}$$

Conditional Likelihood Function

The conditional log likelihood for this model is given below:

$$
\begin{aligned}
l(\mathcal{D}, \theta) &= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - \mathbf{w}\mathbf{x}_n^T)^2 \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}\mathbf{x}_n^T)^2 \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w}^T)^T(\mathbf{y} - \mathbf{X}\mathbf{w}^T)
\end{aligned}
$$

Where $\mathbf{y}$ is a column vector of outputs $\mathbf{y}_n = y_n$ and $\mathbf{X}$ is a matrix where each row is a feature vector $\mathbf{X}_{nd} = \mathbf{x}_{nd}$.

Maximum (Conditional) Likelihood Estimates

- Using some basic results from Matrix calculus, we obtain the following gradient equations:

$$\nabla_{\mathbf{w}} l(\mathcal{D}, \theta) = -\frac{1}{\sigma^2}\left(\mathbf{X}^T\mathbf{X}\mathbf{w}^T - \mathbf{X}^T\mathbf{y}\right) = 0$$

$$\frac{\partial}{\partial\sigma} l(\mathcal{D}, \theta) = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(y_n - \mathbf{w}\mathbf{x}_n^T)^2 = 0$$

- The final MLEs for the parameters $\mathbf{w}_*$ and σ_* are given below:

$$\mathbf{w}_*^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\sigma_* = \left(\frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{w}\mathbf{x}_n^T)^2\right)^{1/2}$$

Prediction

- We previously derived that the optimal prediction under the squared loss function is $f_*(x) = \mathbb{E}_{P_*(Y|\mathbf{x})}[y]$.
- We can now plug in $P(Y|\mathbf{x}, \theta_*)$ as an approximation to $P_*(Y|\mathbf{x})$. The resulting prediction function is:

$$f_*(x) \approx \mathbb{E}_{P(Y|x,\theta_*)}[y] = \mathbf{w}_* \mathbf{x}^T$$

- Note that this prediction rule is actually independent of the output noise variance parameter σ, showing that if all we want to do is make predictions, we do not need to learn this parameter.

Properties of MLE: Consistency

- Recall that a parametric probabilistic model $P(Y|\mathbf{x},\theta)$ should be thought of as a set of probability distributions indexed by the parameter $\theta \in \Theta$.
- If there exists a θ_* such that $P_*(Y|\mathbf{x}) = P(Y|\mathbf{x},\theta_*)$ for all $\mathbf{x} \in \mathcal{X}$, then the model is said to be well-specified and Maximum Likelihood Estimation has some special properties.
- First, assume θ_N is the unique MLE of θ found using a data set $\mathcal{D}$ of size N. Then under mild regularity conditions, $\lim_{N\to\infty} P(\|\theta_N - \theta_*\| \geq \epsilon) = 0$ for any $\epsilon > 0$. This property is called consistency.

Properties of MLE: Discussion

- One of the major reasons to use the MLE is that the estimator is consistent. There is also a sense in which the MLE is optimally efficient in terms of the rate at which θ_N converges to θ_* as N increases.
- This basically means that θ_N has a better chance of being closer to θ_* for large N than other estimators.
- However, the theory for MLE breaks down both when N is not large and when the model is misspecified, that is to say, there is no θ_* for which $P_*(Y|\mathbf{x}) = P(Y|\mathbf{x}, \theta_*)$.
- When N is small, other estimators may significantly out-perform the MLE.

Optimality Properties of MLE: Discussion

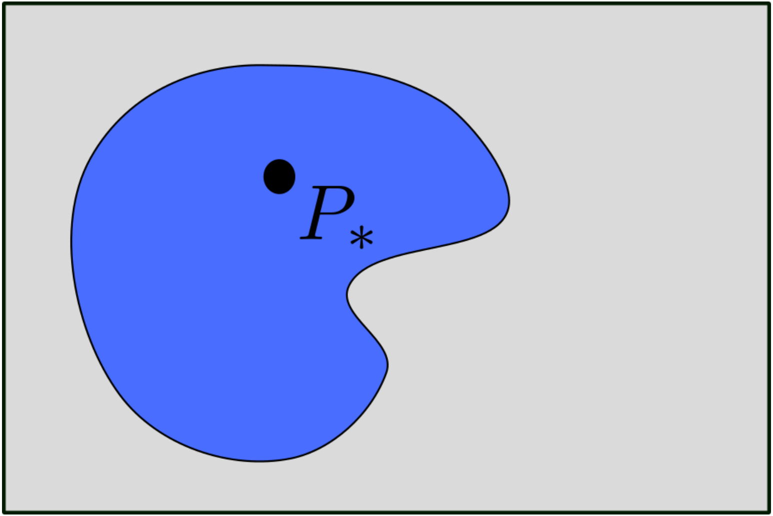
- When the model is misspecified, but the MLE is still unique, it can be shown that MLE finds the distribution within the model that is as “close" as possible to P_* in an expected KL-divergence sense:

$$\theta_* = \min_{\theta} \mathbb{E}_{P_*(X)}[KL(P_*(Y|\mathbf{x})||P(Y|\mathbf{x},\theta))]$$

- Note that KL-divergence is a pre-metric:

$$KL(P(Y)||Q(Y)) = \mathbb{E}_{P(Y)}\left[\log\left(\frac{P(Y)}{Q(Y)}\right)\right]$$

MLE and KL Divergence



MLE and KL Divergence