3. Maximum Likelihood

Varun Kanade
University of Oxford
January 27, 2016
Outline

Probabilistic Framework

- Formulate linear regression in the language of probability
- Introduce the maximum likelihood estimate
- Relation to least squares estimate

Basics of Probability

- Univariate and multivariate normal distribution
- Laplace distribution
- Likelihood, Entropy and its relation to learning
Univariate Gaussian (Normal) Distribution

The univariate normal distribution is defined by the following density function

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ X \sim N(\mu, \sigma^2) \]

Here \( \mu \) is the mean and \( \sigma^2 \) is the variance.

[Diagram showing a normal distribution curve with labeled mean (\( \mu \)) and standard deviation (\( \sigma \)).]

Recall:

\[ \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
\[ \int_{-\infty}^{\infty} x p(x) \, dx = \mu \]
\[ \int_{-\infty}^{\infty} x^2 p(x) \, dx = \mu^2 + \sigma^2 \]
Sampling from a Gaussian distribution

Sampling from $X \sim \mathcal{N}(\mu, \sigma^2)$

By setting $Y = \frac{X - \mu}{\sigma}$, sample from $Y \sim \mathcal{N}(0, 1)$

Cumulative distribution function

$$
\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt
$$
Covariance and Correlation

For random variable $X$ and $Y$ the covariance measures how the random variable change jointly.

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Covariance depends on the scale of the random variable. The (Pearson) correlation coefficient normalizes the covariance to give a value between $-1$ and $+1$.

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$ and $\sigma_Y^2 = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$. 
Multivariate Gaussian Distribution

Suppose $x$ is a $n$-dimensional random vector. The covariance matrix consists of all pairwise covariances.

\[
\text{cov}(x) = \mathbb{E} \left[ (x - \mathbb{E}[x])(x - \mathbb{E}[x])^T \right] = 
\begin{bmatrix}
\text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\
\text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n, X_n)
\end{bmatrix}.
\]

If $\mu = \mathbb{E}[x]$ and $\Sigma = \text{cov}[x]$, the multivariate normal is defined by the density

\[
\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\]

$\Sigma$ is positive definite if $\forall y \neq 0 \in \mathbb{R}^n, y^T \Sigma y > 0$.
Bivariate Gaussian Distribution

Suppose \( X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \) and \( X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \)

What is the joint probability distribution \( p(x_1, x_2) \)?

\[
p(x_1, x_2) = p(x_1) \cdot p(x_2) \quad \text{(if } x_1 \text{ and } x_2 \text{ are independent)}
\]

\[
= \frac{1}{\sqrt{2\pi \sigma_1}} \cdot \frac{1}{\sqrt{2\pi \sigma_2}} \cdot 
\exp \left( -\frac{1}{2} \left( \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right) \right)
\]

\[
\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}
\]

\[
= \left( \frac{1}{(2\pi)^{1/2} |\Sigma|^{1/2}} \right) \exp \left( -\frac{1}{2} \left( x_1 - \mu_1 \right)^T \Sigma^{-1} \left( x_1 - \mu_1 \right) \right)
\]

General case, by rotating (aX + bY)

**The locus of points having same prob density are ellipses around \( \mu \).**

\[
\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = \text{Constant}
\]

**Axis-Aligned**
Suppose you are given three independent samples: 
\(x_1 = 1, \ x_2 = 2.7, \ x_3 = 3.\)

You know that the data were generated from \(\mathcal{N}(0, 1)\) or \(\mathcal{N}(4, 1)\).

Let \(\theta\) represent the parameters of the distribution. Then the probability of observing data with parameter \(\theta\) is called the likelihood:

\[
p(x_1, x_2, x_3 \mid \theta) = p(x_1 \mid \theta)p(x_2 \mid \theta)p(x_3 \mid \theta)
\]

We have to choose between \(\theta = 0\) and \(\theta = 4\). Which one?

\[\mathcal{N}(0, 4) \text{ is more likely}\]

Maximum Likelihood Estimation (MLE): Pick \(\theta\) that maximizes the likelihood.
Recall our linear regression model

\[ y = x^T w + \text{noise} \]

Model \( y \) (conditioned on \( x \)) as a random variable. Given \( x \) and the model parameter \( w \):

\[ \mathbb{E}[y \mid x, w] = x^T w \]

We can be more specific in choosing our model for \( y \). Let us assume that given \( x, w \), \( y \) is Gaussian with mean \( x^T w \) and variance \( \sigma^2 \).

\[ y \sim \mathcal{N}(x^T w, \sigma^2) = x^T w + \mathcal{N}(0, \sigma^2) \]
Likelihood of Linear Regression

Suppose we observe data $(x_i, y_i)_{i=1}^m$. What is the likelihood of observing the data for model parameters $w, \sigma$?
Likelihood of Linear Regression

Suppose we observe data $\langle (x_i, y_i) \rangle_{i=1}^{m}$. What is the likelihood of observing the data for model parameters $w, \sigma$?

$$p (y_1, \ldots, y_m \mid x_1, \ldots, x_m, w, \sigma) = \prod_{i=1}^{m} p (y_i \mid x_i, w, \sigma)$$

Recall that $y_i \sim x_i^T w + \mathcal{N}(0, \sigma^2)$. So

$$p (y_1, \ldots, y_m \mid x_1, \ldots, x_m, w, \sigma) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i^T w)^2}{2\sigma^2}} = \left( \frac{1}{2\pi\sigma^2} \right)^{m/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i-x_i^T w)^2}$$

Want to find parameters $w$ and $\sigma$ that maximize the likelihood
Likelihood of Linear Regression

It is simpler to look at the log-likelihood. Taking logs

\[
LL(y_1, \ldots, y_m \mid x_1, \ldots, x_m, w, \sigma) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i^T w - y_i)^2
\]

\[
LL(y \mid X, w, \sigma) = -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Xw - y)^T (Xw - y)
\]

How to find \(w\) that maximizes the likelihood?

\[
\nabla_w LL = -\frac{1}{2\sigma^2} (2X^T Xw - 2X^T y) = 0
\]

\[
w = (X^T X)^{-1} X^T y \quad \text{(same as least square estimate)}
\]

Make sure that Hessian is negative definite to be sure that you have a maximum.
Maximum Likelihood and Least Squares

Let us in fact look at negative log-likelihood (which is more like loss)

$$NLL(y | X, w, \sigma) = \frac{1}{2\sigma^2} (Xw - y)^T (Xw - y) + \frac{m}{2} \log(2\pi\sigma^2)$$

And recall the squared loss objective

$$L(w) = (Xw - y)^T (Xw - y)$$

We can also find the MLE for $\sigma$. As exercise show that the MLE of $\sigma$ is

$$\sigma_{ML}^2 = \frac{1}{m} (Xw_{ML} - y)^T (Xw_{ML} - y)$$
Making Prediction

Given training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ we can use MLE estimators to make predictions on new points and also give confidence intervals.

$$y_{\text{new}} \mid x_{\text{new}}, \mathcal{D} \sim \mathcal{N}(x_{\text{new}}^T \hat{w}_{\text{ML}}, \sigma_{\text{ML}}^2)$$
Outliers and Laplace Distribution

With outliers least squares (and hence MLE with Gaussian model) can be quite bad.

Instead, we can model the noise (or uncertainty) in $y$ as a Laplace distribution

$$p(y | x, w, b) = \frac{1}{2b} \exp \left( -\frac{|y - x^T w|}{b} \right)$$
Bernoulli random variable $X$ takes value in $\{0, 1\}$. We parametrize using $\theta \in [0, 1]$.

$$p(1 \mid \theta) = \theta$$
$$p(0 \mid \theta) = 1 - \theta$$

More succinctly, we can write

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

For classification, we will design models with parameter $w$ that given input $x$ produce a value in $f(x; w) \in [0, 1]$. Then, we can model the (binary) class labels as:

$$y \sim \text{Bernoulli}(f(x; w))$$
In information theory, entropy $H$ is a measure of uncertainty associated with a random variable.

$$H(X) = -\sum_x p(x) \log(p(x))$$

In the case of bernoulli variables (with parameter $\theta$) we get:

$$H(X) = -\theta \log(\theta) - (1 - \theta) \log(1 - \theta)$$
Suppose we get data $x_1, \ldots, x_m$ from some unknown distribution $q$.

Attempt to find parameters $\theta$ for a family of distributions that best explains the data

$$
\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{m} p(x_i | \theta)
$$

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

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

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

$$
\rightarrow \arg\min_{\theta} \int \log \left( \frac{q(x)}{p(x)} \right) q(x) dx = \text{KL}(q||p)
$$
Kullback-Leibler Divergence

KL-Divergence is “like” a distance between distributions

$$KL(q \| p) = \sum_i \log \frac{q(x_i)}{p(x_i)} q(x_i) dx$$

$$KL(q \| q) = 0$$

$$KL(q \| p) \geq 0 \text{ for all distributions } p$$