Parameter Estimation, Prediction and Inference

We have finished with discussing the basic core ideas in probability. Congratulations!

In the remaining of the course, we will discuss some applications of probability to some key problems in statistics, machine learning and signal processing. While our earlier examples focus on showing how to use probability to calculate various quantities of interest, in these applications, we will show how probability can be used to derive widely-used algorithms for various tasks. In particular, we will talk about the Viterbi algorithm and the EM algorithm.

Like for all probability problems, the center of action is the probability model, with a set of parameters \( \theta_1, \theta_2, \ldots, \theta_k \). There are usually two key steps in the use of this model in these problems:

- First, the parameters need to be estimated from data.
- Second, the model is used to perform prediction and inference, i.e. predict unobserved random variables from observed ones.

We have already seen examples of both tasks earlier in the course. Polling to estimate the fraction of populations who would vote for a Democrat is an example of the first task; there, the unknown parameter is \( p \), the fraction of the population who will vote for a Democrat. Inferring the disease state of a patient from the result of a medical test is an example of the second task. In the next few lectures, we will discuss more complex examples where the end results of our analysis will be algorithms rather than just a calculation. Moreover, we will discuss an application from speech recognition where the two tasks are intertwined.

Let us first discuss how to formulate the problem of parameter estimation in general.

Maximum Likelihood Estimation

Recall the polling example. A natural estimator for \( p \), the fraction of the population who will vote for a Democrat, is:

\[
\hat{p} = \frac{\text{number of people polled who will vote for a Democrat}}{\text{number of people polled}}.
\]

This estimator is an intuitive one for this simple problem, but can be derived from a general principle called maximum likelihood (ML) principle: we are actually choosing as an estimate the value of \( p \) that maximizes the probability, or the **likelihood**, of observing the data:

\[
\hat{p}_{ML} = \arg\max_p P(\text{observation}; p),
\]

where the observation here is the sequence of answers from the people polled. Why? \( P(\text{observation}; p) = p^k(1-p)^{n-k} \), where \( n \) is the total number of people polled, and \( k \) is the number or people who will vote for
a Democrat. By simple calculus,
\[
\arg\max_p p^k (1 - p)^{n-k} = \frac{k}{n},
\]
which is the intuitive estimate we came up with earlier.
We will discuss a few more examples of the ML principle below.

**Maximum Likelihood for Gaussian observations**

Suppose we observe \( n \) i.i.d. samples \( Y_1, \ldots, Y_n \) such that \( Y_i \sim N(\mu, \sigma^2) \), where \( \sigma^2 \) is known but \( \mu \) is unknown. What would be a reasonable estimate of \( \mu \)? Another equivalent way to interpret this model is that the \( Y_i \)'s are all noisy observations of an unknown signal \( \mu \):
\[
Y_i = \mu + Z_i,
\]
where \( Z_i \)'s are i.i.d. \( N(0, \sigma^2) \) noise corrupting each of the observations. Our goal is to estimate the unknown signal from the observations.

In the case of discrete-valued observations, the maximum likelihood principle chooses the value of the parameters to maximizes the probability of the observation. But for continuous-valued observations, like for this Gaussian problem, the probability of any observation is zero! So instead, we replace the probability by the density and use that to define the likelihood. So, the likelihood for this problem is:
\[
f(y_1, \ldots, y_n; \mu) = \prod_{i=1}^n f(y_i; \mu) \\
= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu)^2}{2\sigma^2} \right).
\]
It is the product of the pdf’s evaluated at the samples.

Since the likelihood involves a product, it is sometimes more convenient to work with the log-likelihood
\[
\log f(y_1, \ldots, y_n; \mu) = \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu)^2}{2\sigma^2} \right) \right) \\
= -n \log \left( \sqrt{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.
\]
This is a quadratic function of \( \mu \), with a maximum at the point where the derivative of the function is zero. By differentiating this function with respect to \( \mu \) and setting the derivative to zero and solving for \( \mu \), we obtain:
\[
\mu^\star = \frac{1}{n} \sum_{i=1}^n y_i.
\]
Hence, the maximum likelihood estimator for this problem is:
\[
\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n Y_i.
\]
Linear Regression

Suppose we have a dataset on \( n \) houses with sizes \( x_1, \ldots, x_n \) square feet and prices \( y_1, \ldots, y_n \) dollars. We would like to model the dependency of \( y_i \) on \( x_i \). In the model of linear regression, we assume the data comes from a probability model:

\[
Y_i = aX_i + b + Z_i, \quad Z_i \sim N(0, \sigma^2),
\]

where we assume the \( X_i \)'s and \( Z_i \)'s are mutually independent random variables.

The parameter we would like to estimate is \( \theta = (a, b) \). This is equivalent to finding the slope and the y-intercept of a line that fits the data points. After we estimate the parameters by the maximum likelihood estimate \( \hat{a}_{\text{ML}}, \hat{b}_{\text{ML}} \), the predictor for the data point \( x \) would be \( \hat{y} = \hat{a}_{\text{ML}}x + \hat{b}_{\text{ML}} \). Note that \( \sigma^2 \) is also an unknown in the model that we can estimate. However, since \( \sigma^2 \) does not affect the predictor \( \hat{y} = \hat{a}_{\text{ML}}x + \hat{b}_{\text{ML}} \), we focus on estimating \( a \) and \( b \) here.

Note that the problem we considered in the last lecture of estimating the mean \( \mu \) of \( n \) i.i.d. observations \( Y_1, Y_2, \ldots, Y_n \) is a special case of the linear regression problem, with \( a = 0 \) and \( b = \mu \). However, in typical applications of statistics and machine learning, we observe features (house size in our example) and want to use them to predict a target variable of interest (house price in our example). So the more general linear regression model with \( a \neq 0 \) is of much more interest.

To find the maximum likelihood estimate of the parameters, we find \( (a, b) \) which maximizes the joint density \( f(x_1, y_1, \ldots, x_n, y_n; a, b) \). Since we assume each data point is independent,

\[
f(x_1, y_1, \ldots, x_n, y_n; a, b) = f(x_1, y_1; a, b) \cdots f(x_n, y_n; a, b).
\]

Consider the individual terms \( f(x_i, y_i; a, b) \). Since the distribution of \( X_i \) does not depend on \( a, b \),

\[
f(x_i, y_i; a, b) = f(x_i)f(y_i|x_i; a, b)
\]

\[= f(x_i) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right).
\]

Hence finding the MLE is equivalent to minimizing over all \( a \) and \( b \):

\[
\sum_{i=1}^{n}(y_i - ax_i - b)^2.
\]

This is called the least squares problem. We are fitting a line through the data points that minimizes the sum of the squared errors. The minimizing \( (a, b) \) is given as

\[
\begin{bmatrix}
  a^* \\
  b^*
\end{bmatrix} = (A^TA)^{-1}A^T
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix}, \quad \text{where } A = \begin{bmatrix}
  x_1 & x_2 & \cdots & x_n \\
  1 & 1 & \cdots & 1
\end{bmatrix}^T.
\]

The above fact can be proved by differentiating the expression wrt \( a \) and \( b \) and setting the derivatives to 0. For this, we note that

\[
\sum_{i=1}^{n}(y_i - ax_i - b)^2 = ||y - A\begin{bmatrix}a \\
b\end{bmatrix}||^2
\]

where

\[
y = \begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix}
\]
and $\|z\|^2$ of a vector $z$ is the sum of squares of the entries in $z$. The gradient of this wrt $\begin{bmatrix} a \\ b \end{bmatrix}$ is $A^T(y - A \begin{bmatrix} a \\ b \end{bmatrix})$. On setting this to zero, we get the desired result.

Note that the probabilistic model with Gaussian additive noise justifies the squared error as a criterion of measuring the goodness of fit. If the model is different, then the criterion would be different. However, maximum likelihood parameter estimation is a general principle that can be applied to any probabilistic model of how the data is generated and allows us to derive the appropriate data fitting procedure.