Probabilistic perspective of Machine Learning

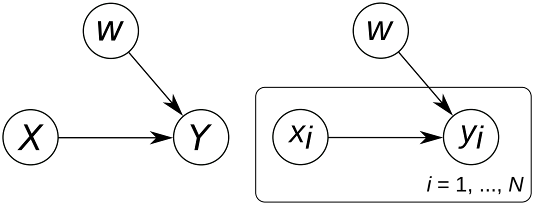
András Attila Sulyok

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics

2024-04-09

Probabilistic model

Supervised learning:



Likelihood (of the parameters): $P(\mathcal{D} \mid w) = P(Y \mid X, w)P(X)$

1. Probabilistic perspective
2. Maximum Likelihood Estimation
3. Maximum A Posteriori Estimation

Maximum Likelihood Estimation

$$\arg\max_{w} P(\mathcal{D} \mid w) = \arg\max_{w} \{P(Y \mid X, w)P(X)\}$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\begin{aligned}\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \{P(Y \mid X, w)P(X)\} \\ &= \arg\max_w P(Y \mid X, w)\end{aligned}$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\begin{aligned}\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \{P(Y \mid X, w)P(X)\} \\ &= \arg\max_w P(Y \mid X, w) \\ &= \arg\max_w P(y_1, \ldots y_N \mid x_1, \ldots, x_N, w)\end{aligned}$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\begin{aligned}
\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \{P(Y \mid X, w) P(X)\} \\
&= \arg\max_w P(Y \mid X, w) \\
&= \arg\max_w P(y_1, \ldots y_N \mid x_1, \ldots, x_N, w) \\
&= \arg\max_w \prod_{i=1}^{N} P(y_i \mid x_i, w) \qquad\qquad y\text{s are i.i.d}
\end{aligned}$$

Assumption: samples are independent and indentically distributed.

Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$
\begin{aligned}
\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \{P(Y \mid X, w)P(X)\} \\
&= \arg\max_w P(Y \mid X, w) \\
&= \arg\max_w P(y_1, \ldots y_N \mid x_1, \ldots, x_N, w) \\
&= \arg\max_w \prod_{i=1}^{N} P(y_i \mid x_i, w) \qquad\qquad y\text{s are i.i.d}
\end{aligned}
$$

(some algebraic magic follows)

$$
= \arg\max_w \log \prod_{i=1}^{N} P(y_i \mid x_i, w)
$$

Assumption: samples are independent and indentically distributed.

Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\arg\max_w P(\mathcal{D} \mid w) = \arg\max_w \log \prod_{i=1}^{N} P(y_i \mid x_i, w)$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\begin{aligned}\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \log \prod_{i=1}^{N} P(y_i \mid x_i, w) \\ &= \arg\max_w \sum_{i=1}^{N} \log P(y_i \mid x_i, w)\end{aligned}$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$\begin{aligned}
\arg\max_{w} P(\mathcal{D} \mid w) &= \arg\max_{w} \log \prod_{i=1}^{N} P(y_i \mid x_i, w) \\
&= \arg\max_{w} \sum_{i=1}^{N} \log P(y_i \mid x_i, w) \\
&= \arg\min_{w} \sum_{i=1}^{N} - \log P(y_i \mid x_i, w)
\end{aligned}$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Maximum Likelihood Estimation

$$
\begin{aligned}
\arg\max_w P(\mathcal{D} \mid w) &= \arg\max_w \log \prod_{i=1}^{N} P(y_i \mid x_i, w) \\
&= \arg\max_w \sum_{i=1}^{N} \log P(y_i \mid x_i, w) \\
&= \arg\min_w \sum_{i=1}^{N} - \log P(y_i \mid x_i, w) \\
&= \boxed{\arg\min_w \mathbb{E}_{(x,y)\sim\mathcal{D}} - \log P(y \mid x, w)}
\end{aligned}
$$

Assumption: samples are independent and indentically distributed.
Assumption: distribution of x is uniform

Example: logistic regression

Remember the likelihood for the logistic regression:

$$P(y = + \mid x, w) = \theta(w^T x)$$

(θ is the logistic sigmoid)

Example: logistic regression

Remember the likelihood for the logistic regression:

$$P(y = + \mid x, w) = \theta(w^T x)$$

(θ is the logistic sigmoid)

Binary cross entropy error (for one sample point):

$$\begin{cases} -\log(1 - \theta(w^T x)) & \text{if } y = - \\ -\log \theta(w^T x) & \text{if } y = + \end{cases}$$

Example: linear regression

Likelihood:

$$P(y \mid x, w) = \mathcal{N}(y; w^T x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y - w^T x\right)^2}{2\sigma^2}\right)$$

Example: linear regression

Likelihood:

$$P(y \mid x, w) = \mathcal{N}(y; w^T x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)$$

Cross entropy error for linear regression:

$$\begin{aligned}\arg\min_w -\log P(y \mid x, w) &= \arg\min_w \frac{1}{2}\log\left(2\pi\sigma^2\right) + \left(\frac{(y - w^T x)^2}{2\sigma^2}\right) \\ &= \arg\min_w \left(y - w^T x\right)^2\end{aligned}$$

Looks familiar?

Maximum A Posteriori Estimation

ML objective was:

$$\arg\max_{w} P(\mathcal{D} \mid w)$$

Maximum A Posteriori Estimation

ML objective was:

$$\arg\max_{w} P(\mathcal{D} \mid w)$$

Alternative objective:

$$\arg\max_{w} P(w \mid \mathcal{D})$$

Maximum A Posteriori Estimation

ML objective was:

$$\arg\max_w P(\mathcal{D} \mid w)$$

Alternative objective:

$$\arg\max_w P(w \mid \mathcal{D}) = \arg\max_w \frac{P(\mathcal{D} \mid w) P(w)}{P(\mathcal{D})}$$

Maximum A Posteriori Estimation

ML objective was:

$$\arg\max_w P(\mathcal{D} \mid w)$$

Alternative objective:

$$\arg\max_w P(w \mid \mathcal{D}) = \arg\max_w \frac{P(\mathcal{D} \mid w)P(w)}{P(\mathcal{D})} = \arg\max_w P(\mathcal{D} \mid w)P(w)$$

Similar to before, but there is a prior:
incorporates prior knowledge

Example: linear regression

Exercise 1 :)

Aside: Bayesian inference

During training, calculate:

$$P(w \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid w)P(w)}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid w)P(w)}{\int_{\mathbb{R}^d} P(\mathcal{D} \mid w')P(w')dw}$$

The evidence is usually hard to calculate

Aside: Bayesian inference

During training, calculate:

$$P(w \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid w)P(w)}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid w)P(w)}{\int_{\mathbb{R}^d} P(\mathcal{D} \mid w')P(w')dw}$$

The evidence is usually hard to calculate

Inference (prediction) for a new sample point x':

$$P(y \mid x') = \int_{\mathbb{R}^d} P(y \mid x', w)P(w \mid \mathcal{D})dw$$

It is not a point estimate:
it gives a distribution over possible parameter/prediction values.