COMPSCI 689
Lecture 14: Probabilistic Supervised learning

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

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

Parametric Probability Mass Function

Definition: A parametric probability mass function $P : \mathcal{Z} \rightarrow \mathbb{R}$ for a discrete random variable Z is a function that satisfies the requirements below for any value of the parameters ϕ with parameter space Φ :

- Normalization: $\sum_{z \in \mathcal{Z}} P(Z = z|\phi) = 1$
- Non-Negativity: $\forall z \in \mathcal{Z} \; P(Z = z|\phi) \geq 0$

Example: Bernoulli Random Variables

Consider a biased coin. Let ϕ be the probability that the coin comes up heads. Let $Z \in \{0, 1\}$ be the value of the coin flip (1 for heads, 0 for tails). We thus have:

- Values: $\mathcal{Z} = \{0, 1\}$
- Parameters: $\phi \in [0, 1]$
- Parameter Space: $\Phi = [0, 1]$
- Mass Function: $P(Z = z|\phi) = \phi^z(1 - \phi)^{(1-z)}$

Parametric Probability Density Function

Definition: A parametric probability density function $p : \mathcal{Z} \rightarrow \mathbb{R}$ for a continuous random variable Z is a function that satisfies the requirements below for any value of the parameters ϕ with parameter space Φ :

- Normalization: $\int_{\mathcal{Z}} p(Z = z|\phi)dz = 1$
- Non-Negativity: $\forall z \in \mathcal{Z}\ p(Z = z|\phi) \geq 0$

Example: Normal Random Variables

Consider the univariate normal random variable Z.

- Values: $\mathcal{Z} = \mathbb{R}$
- Parameters: $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^{>0}$
- Parameter Space: $\Phi = \mathbb{R} \times \mathbb{R}^{>0}$
- Density Function: $p(Z = z|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(z - \mu)^2)$

Probabilistic Models

- Definition: A parametric probability model $\mathbb{P}$ for discrete random variable Z with parametric probability mass function P and parameter space Φ is the set of all probability mass functions generated by P and Φ:

$$\mathbb{P} = \{P(Z|\phi)|\phi \in \Phi\}$$

- Definition: A parametric probability model $\mathbb{P}$ for a continuous random variable Z with with parametric probability density function P and parameter space Φ is the set of all probability density functions generated by p and Φ:

$$\mathbb{P} = \{p(Z|\phi)|\phi \in \Phi\}$$

Parameter Estimation for Probabilistic Models

Let $\mathcal{L}(\mathcal{D}, \phi)$ be a loss function that measures how “close" a data set $\mathcal{D}$ is to a parametric probability mass or density function with parameters ϕ. Given such a loss function, we can estimate the parameters as follows.

The Parameter Estimation Problem

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \mathcal{L}(\mathcal{D}, \phi)$$

This selects the best fitting distribution from the model $\mathbb{P}$.

Maximum Likelihood Estimation

- The most common such loss function used to estimate the parameters of probabilistic models is the negative log likelihood function:

$$nll(\mathcal{D}, \phi) = -\sum_{n=1}^{N} \log P(Z = z_n|\phi)$$

- This loss function derives from the idea of selecting the parameters ϕ that make the data the most likely:

$$\hat{\phi} = \arg\max_{\phi \in \Phi} \prod_{n=1}^{N} P(Z = z_n|\phi)$$

Example: Bernoulli Distribution

- Let $P(Z = z|\phi) = \phi^z(1-\phi)^{(1-z)}$ for $\mathcal{Z} = \{0, 1\}$ and $\phi \in [0, 1]$.
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of ϕ.
- The negative log likelihood function is shown below and is subject to $\phi \in [0, 1]$:

$$nll(\mathcal{D}, \phi) = -\sum_{n=1}^{N}(z_n \log\phi + (1-z_n)\log(1-\phi))$$

- The MLE is $\hat{\phi} = \frac{1}{N}\sum_{n=1}^{N} z_n$.

Example: Normal Mean

- Let $p(Z = z|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right)$
- Suppose we have a data set $\mathcal{D} = [z_1, ..., z_N]$ and we want to find the MLE of μ.

- The negative log likelihood function is:

$$nll(\mathcal{D}, \mu, \sigma) = \sum_{n=1}^{N} \left(\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(z_n - \mu)^2 \right)$$

- The MLE is $\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} z_n$.

Probabilistic Supervised Learning

- In probabilistic supervised learning, our goal is to model the true probability distribution of the outputs $y \in \mathcal{Y}$ given the inputs $\mathbf{x} \in \mathcal{X}$.
- If $\mathcal{Y}$ is discrete, our goal is to model $P_*(Y = y|\mathbf{X} = \mathbf{x})$ with a conditional parametric probability mass function $P(Y = y|\mathbf{X} = \mathbf{x}, \theta)$.
- If $\mathcal{Y}$ is uncountable, our goal is to model $p_*(Y = y|\mathbf{X} = \mathbf{x})$ with a conditional parametric probability density function $p(Y = y|\mathbf{X} = \mathbf{x}, \theta)$.

Generating Supervised Probabilistic Models

- We can very flexibly generate probabilistic supervised learning models by combining unconditional probability models with regression models that predict their parameter values.
- We can select different probability models to provide distributions over different output spaces.
- We can use any type of regression model including both linear and non-linear models.
- We may need to apply a transformation to the regression model outputs to ensure that the predicted parameter values always fall in the parameter space of the unconditional model.

Learning Supervised Probabilistic Models

- So long as all model components are differentiable functions, we can learn the model parameters θ by minimizing the conditional negative log likelihood function given a data set $\mathcal{D}$:

$$nll(\mathcal{D},\theta) = -\sum_{n=1}^{N} \log P(Y = y|\mathbf{X} = \mathbf{x}, \theta)$$

- This is technically maximum conditional likelihood estimation, but is also often just referred to as maximum likelihood estimation.

Example: Probabilistic Linear Regression

- Suppose that $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} \in \mathbb{R}^D$.
- Unconditional Model:
  $p(Y = y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$
- Conditional Model:

$$p(Y = y|\mathbf{X} = \mathbf{x}, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - f_w(\mathbf{x}))^2\right)$$

- Parameter Prediction Function: $f_w(\mathbf{x}) = \mathbf{x}\mathbf{w}$
- Model Parameters: $\theta = [\mathbf{w}, \sigma]$
- Negative Log Likelihood:

$$nll(\mathcal{D}, \theta) = -\sum_{n=1}^{N} \log p(Y = y_n|\mathbf{X} = \mathbf{x}_n, \theta)$$

Example: Probabilistic Linear Regression

$$
\begin{aligned}
nll(\mathcal{D}, \theta) &= -\sum_{n=1}^{N} \log p(Y = y_n | \mathbf{X} = \mathbf{x}_n, \theta) \\
&= -\sum_{n=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_n - f_w(\mathbf{x}_n))^2 \right) \right) \\
&= -\sum_{n=1}^{N} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_n - \mathbf{x}_n \mathbf{w})^2 \right) \\
&= \sum_{n=1}^{N} \left(\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} (y_n - \mathbf{x}_n \mathbf{w})^2 \right)
\end{aligned}
$$

Minimizing this NLL for a fixed value of σ is equivalent to learning the OLS linear regression model under ERM.

Example: Probabilistic Non-linear Regression

- Suppose that $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} \in \mathbb{R}^D$.
- Unconditional Model:
  $p(Y = y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$
- Conditional Model:

$$p(Y = y|\mathbf{X} = \mathbf{x}, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - f(\mathbf{x}))^2\right)$$

- Parameter Prediction Function: $f(\mathbf{x})$ = an MLP, CNN, RNN, transformer, etc.
- Negative Log Likelihood:

$$nll(\mathcal{D}, \theta) = -\sum_{n=1}^{N} \log p(Y = y_n|\mathbf{X} = \mathbf{x}_n, \theta)$$

Making Predictions

- Given a probabilistic supervised model, we can produce an estimate of the conditional probability of y given $\mathbf{x}$ by plugging the estimated parameters $\hat{\theta}$ into the model.
- In the case of discrete y, when we need to issue a prediction, we typically predict the value that achieves the maximum conditional probability given $\mathbf{x}$ and $\hat{\theta}$:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(Y = y | \mathbf{X} = \mathbf{x}, \hat{\theta})$$

- In the case of continuous y, we can predict different functions of the conditional distribution. The most commonly used prediction is the conditional mean of y:

$$\hat{y} = E_{p(Y=y|\mathbf{X}=\mathbf{x},\hat{\theta})}[y] = \int_{\mathcal{Y}} yp(Y = y|\mathbf{X} = \mathbf{x}, \hat{\theta})dy$$