Machine Learning (CSE 446): Probabilistic View of Logistic Regression and Linear Regression

Noah Smith

University of Washington
nasmith@cs.washington.edu

October 30, 2017

Remember the Bayes optimal classifier. $\mathcal{D}$ is the true probability distribution over input-output pairs.

$$f^{(\mathsf{BO})}(x) = \operatorname*{argmax}_y \mathcal{D}(x, y)$$

Remember the Bayes optimal classifier. $\mathcal{D}$ is the true probability distribution over input-output pairs.

$$
\begin{aligned}
f^{(\mathsf{BO})}(x) &= \operatorname*{argmax}_y \mathcal{D}(x, y) \\
&= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \cdot \mathcal{D}(x)
\end{aligned}
$$

Remember the Bayes optimal classifier. $\mathcal{D}$ is the true probability distribution over input-output pairs.

$$\begin{aligned} f^{(\mathsf{BO})}(x) &= \operatorname*{argmax}_y \mathcal{D}(x, y) \\ &= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \cdot \mathcal{D}(x) \\ &= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \end{aligned}$$

Remember the Bayes optimal classifier. $\mathcal{D}$ is the true probability distribution over input-output pairs.

$$
\begin{aligned}
f^{(\mathsf{BO})}(x) &= \operatorname*{argmax}_y \mathcal{D}(x, y) \\
&= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \cdot \mathcal{D}(x) \\
&= \operatorname*{argmax}_y \mathcal{D}(y \mid x)
\end{aligned}
$$

Of course, we don't have $\mathcal{D}(y \mid x)$.

Remember the Bayes optimal classifier. $\mathcal{D}$ is the true probability distribution over input-output pairs.

$$\begin{aligned} f^{(\mathsf{BO})}(x) &= \operatorname*{argmax}_y \mathcal{D}(x, y) \\ &= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \cdot \mathcal{D}(x) \\ &= \operatorname*{argmax}_y \mathcal{D}(y \mid x) \end{aligned}$$

Of course, we don't have $\mathcal{D}(y \mid x)$.

Probabilistic machine learning: define a probabilistic model relating random variables X and Y, and estimate its parameters.

Logistic Regression as a Probabilistic Model

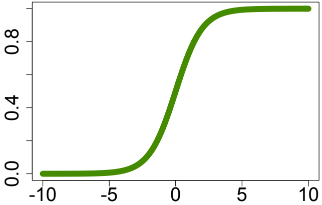
Logistic regression defines $p_{\mathbf{w},b}(Y \mid X)$ as follows:

1. Observe the feature vector $\mathbf{x}$; transform it via the activation function:

$$a = \mathbf{w} \cdot \mathbf{x} + b$$

2. Transform a into a binomial probability by passing it through the logistic function:

$$p_{\mathbf{w},b}(Y = +1 \mid \mathbf{x}) = \frac{1}{1 + \exp -a}$$



3. Sample Y from $p_{\mathbf{w},b}(Y \mid \mathbf{x})$.

Logistic Regression Probabilities

Probability that $Y = +1$ given $\mathbf{x}$:

$$\frac{1}{1 + \exp - (\mathbf{w} \cdot \mathbf{x} + b)}$$
$$= \frac{1}{1 + \exp -y (\mathbf{w} \cdot \mathbf{x} + b)}$$

Approaches 1 as $\mathbf{w} \cdot \mathbf{x} + b \to +\infty$.
Never gets to 0.

Probability that $Y = -1$ given $\mathbf{x}$:

$$1 - \frac{1}{1 + \exp - (\mathbf{w} \cdot \mathbf{x} + b)}$$
$$= \frac{1}{1 + \exp (\mathbf{w} \cdot \mathbf{x} + b)}$$
$$= \frac{1}{1 + \exp -y (\mathbf{w} \cdot \mathbf{x} + b)}$$

Approaches 1 as $\mathbf{w} \cdot \mathbf{x} + b \to -\infty$.
Never gets to 0.

Maximum Likelihood Estimation

The principle of maximum likelihood estimation is to choose parameters (today, $\mathbf{w}$ and b) that make the training data as likely as possible.

Maximum Likelihood Estimation

The principle of maximum likelihood estimation is to choose parameters (today, $\mathbf{w}$ and b) that make the training data as likely as possible.
Mathematically:

$$(\hat{\mathbf{w}}, \hat{b}) = \operatorname*{argmax}_{\mathbf{w}, b} \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n)$$

Maximum Likelihood Estimation

The principle of maximum likelihood estimation is to choose parameters (today, $\mathbf{w}$ and b) that make the training data as likely as possible.
Mathematically:

$$
\begin{aligned}
(\hat{\mathbf{w}}, \hat{b}) &= \operatorname*{argmax}_{\mathbf{w}, b} \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \operatorname*{argmax}_{\mathbf{w}, b} \log \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n)
\end{aligned}
$$

Maximum Likelihood Estimation

The principle of maximum likelihood estimation is to choose parameters (today, $\mathbf{w}$ and b) that make the training data as likely as possible.
Mathematically:

$$
\begin{aligned}
(\hat{\mathbf{w}}, \hat{b}) &= \operatorname*{argmax}_{\mathbf{w},b} \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \operatorname*{argmax}_{\mathbf{w},b} \log \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \operatorname*{argmax}_{\mathbf{w},b} \sum_{n=1}^{N} \log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n)
\end{aligned}
$$

Maximum Likelihood Estimation

The principle of maximum likelihood estimation is to choose parameters (today, $\mathbf{w}$ and b) that make the training data as likely as possible.
Mathematically:

$$
\begin{aligned}
(\hat{\mathbf{w}}, \hat{b}) &= \underset{\mathbf{w},b}{\operatorname{argmax}} \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \underset{\mathbf{w},b}{\operatorname{argmax}} \log \prod_{n=1}^{N} p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \underset{\mathbf{w},b}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \\
&= \underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{n=1}^{N} -\log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n)
\end{aligned}
$$

Logistic Regression-MLE is (Unregularized) Log Loss Minimization!

$$\operatorname*{argmin}_{\mathbf{w},b} \sum_{n=1}^{N} -\log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \equiv \operatorname*{argmin}_{\mathbf{w},b} \frac{1}{N} \sum_{n=1}^{N} LogLoss_n(\mathbf{w}, b)$$

Linear Regression as a Probabilistic Model

Linear regression defines $p_{\mathbf{w},b}(Y \mid X)$ as follows:

1. Observe the feature vector $\mathbf{x}$; transform it via the activation function:

$$\mu = \mathbf{w} \cdot \mathbf{x} + b$$

2. Let μ be the mean of a normal distribution and define the density:

$$p_{\mathbf{w},b}(Y \mid \mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(Y-\mu)^2}{2\sigma^2}$$

3. Sample Y from $p_{\mathbf{w},b}(Y \mid \mathbf{x})$.

Linear Regression-MLE is (Unregularized) Squared Loss Minimization!

$$\operatorname*{argmin}_{\mathbf{w},b} \sum_{n=1}^{N} -\log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \equiv \operatorname*{argmin}_{\mathbf{w},b} \frac{1}{N} \sum_{n=1}^{N} \underbrace{(y_n - (\mathbf{w} \cdot \mathbf{x}_n + b))^2}_{SquaredLoss_n(\mathbf{w},b)}$$

Linear Regression-MLE is (Unregularized) Squared Loss Minimization!

$$\operatorname*{argmin}_{\mathbf{w},b} \sum_{n=1}^{N} -\log p_{\mathbf{w},b}(y_n \mid \mathbf{x}_n) \equiv \operatorname*{argmin}_{\mathbf{w},b} \frac{1}{N} \sum_{n=1}^{N} \underbrace{(y_n - (\mathbf{w} \cdot \mathbf{x}_n + b))^2}_{SquaredLoss_n(\mathbf{w},b)}$$

Where did the variance go?