Machine Learning (CSE 446): Probabilistic Machine Learning

Sham M Kakade

© 2018

University of Washington
skakade@cs.washington.edu
You can think of MLE as a “black box” for choosing parameter values.
Understanding MLE
Understanding MLE
Understanding MLE

\[ x \times w \times \sum b \times \text{logistic} \rightarrow Y \]

\[ \hat{w}, \hat{b} \rightarrow \text{MLE} \]
Probabilistic Stories

logistic regression

Bernoulli

\[ x \times w \times \sum b \text{logistic} \]

\[ \pi \text{logistic regression} \]

\[ \text{Bernoulli} \]

\[ Y \]
Probabilistic Stories

logistic regression

\[ x \times w \times \sum b \rightarrow \text{logistic} \rightarrow Y \]

Bernoulli

\[ \pi \rightarrow Y \]

Gaussian

\[ \mu \rightarrow Y \]

\[ \sigma^2 \rightarrow Y \]

linear regression

\[ x \times w \times \sum b \rightarrow \text{linear regression} \rightarrow Y \]

\[ \sigma^2 \rightarrow Y \]
Then and Now

Before today, you knew how to do MLE:

- For a Bernoulli distribution: $\hat{\pi} = \frac{\text{count}(+1)}{\text{count}(+1) + \text{count}(-1)} = \frac{N^+}{N}$

- For a Gaussian distribution: $\hat{\mu} = \frac{\sum_{n=1}^{N} y_n}{N}$ (and similar for estimating variance, $\hat{\sigma}^2$).

Logistic regression and linear regression, respectively, generalize these so that the parameter is itself a function of $x$, so that we have a **conditional model** of $Y$ given $X$.

- The practical difference is that the MLE doesn’t have a closed form for these models.
  (So we use SGD and friends.)
There is a closed form for the MLE of linear regression.

To keep it simple, assume $b = 0$.

Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be the stack of training inputs and $\mathbf{y} \in \mathbb{R}^N$ be the stack of training outputs.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} (y_n - \mathbf{w} \cdot \mathbf{x}_n)^2$$
A Twist!

There is a closed form for the MLE of linear regression.

To keep it simple, assume $b = 0$.

Let $X \in \mathbb{R}^{N \times d}$ be the stack of training inputs and $y \in \mathbb{R}^{N}$ be the stack of training outputs.

$$\hat{w} = \arg\min_{w} \frac{1}{N} \sum_{n=1}^{N} (y_n - w \cdot x_n)^2 \equiv \arg\min_{w} (y - Xw)^\top (y - Xw)$$
A Twist!

There is a closed form for the MLE of linear regression.

To keep it simple, assume $b = 0$.

Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be the stack of training inputs and $\mathbf{y} \in \mathbb{R}^N$ be the stack of training outputs.

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} (y_n - \mathbf{w} \cdot \mathbf{x}_n)^2 \equiv \arg \min_{\mathbf{w}} (\mathbf{y} - \mathbf{Xw})^\top (\mathbf{y} - \mathbf{Xw})$$

gradient w.r.t. $\mathbf{w}$

$$-2\mathbf{X}^\top (\mathbf{y} - \mathbf{Xw}) = 0$$
A Twist!

There is a closed form for the MLE of linear regression.

To keep it simple, assume $b = 0$.

Let $X \in \mathbb{R}^{N \times d}$ be the stack of training inputs and $y \in \mathbb{R}^N$ be the stack of training outputs.

\[
\hat{w} = \arg\min_w \frac{1}{N} \sum_{n=1}^{N} (y_n - w \cdot x_n)^2 \equiv \arg\min_w (y - Xw)^\top (y - Xw)
\]

\[
\text{gradient w.r.t. } w = \frac{\partial}{\partial w} \left( -2X^\top (y - Xw) \right) = 0
\]

\[
\hat{w} = \left( X^\top X \right)^{-1} X^\top y
\]

Invertibility is fine if we have more than $d$ linearly independent observations.
A Twist!

There is a closed form for the MLE of linear regression.
To keep it simple, assume $b = 0$.
Let $X \in \mathbb{R}^{N \times d}$ be the stack of training inputs and $y \in \mathbb{R}^N$ be the stack of training outputs.

\[
\hat{w} = \arg\min_w \frac{1}{N} \sum_{n=1}^{N} (y_n - w \cdot x_n)^2 \equiv \arg\min_w (y - Xw)^\top (y - Xw)
\]

\[
\text{gradient w.r.t. } w \\
-2X^\top (y - Xw) = 0
\]

\[
\hat{w} = \left( X^\top X \right)^{-1} X^\top y
\]

Invertibility is fine if we have more than $d$ linearly independent observations. But it costs $O(d^3)$.
MLE is Dangerous

\[
\text{Variance}(\hat{\pi}) = \frac{\pi(1 - \pi)}{N} \quad \text{(Note that } \pi \text{ is the } true \text{ probability that } Y = 1! \text{)}
\]

\[
\text{Variance}(\hat{\mu}) = \frac{\sigma^2}{N} \quad \text{(Note that } \sigma^2 \text{ is the } true \text{ variance of the r.v.}! \text{)}
\]
MLE is Dangerous

\[
\text{Variance}(\hat{\pi}) = \frac{\pi (1 - \pi)}{N}
\]
(Note that \(\pi\) is the true probability that \(Y = 1\)!) 

\[
\text{Variance}(\hat{\mu}) = \frac{\sigma^2}{N}
\]
(Note that \(\sigma^2\) is the true variance of the r.v.!) 

Recall the bias-variance tradeoff.

- Bias/approximation error: if your choice of features and probabilistic model align to reality, MLE is great.
- Variance/estimation error: MLE tends to overfit unless you have a lot of data.
MLE is Dangerous

\[
\text{Variance}(\hat{\pi}) = \frac{\pi(1 - \pi)}{N} \quad \text{(Note that } \pi \text{ is the \textit{true} probability that } Y = 1!) \]

\[
\text{Variance}(\hat{\mu}) = \frac{\sigma^2}{N} \quad \text{(Note that } \sigma^2 \text{ is the \textit{true} variance of the r.v.!)}
\]

Regularization reduces variance but increases bias.
Adding Regularization to the Probabilistic Story

Probabilistic story:

- For \( n \in \{1, \ldots, N\} \):
  - Observe \( x_n \).
  - Transform it using parameters \( w \) and \( b \) to get \( p_{w,b}(Y \mid x_n) \).
  - Sample \( y_n \sim p_{w,b}(Y \mid x_n) \).
Adding Regularization to the Probabilistic Story

Probabilistic story:

- For $n \in \{1, \ldots, N\}$:
  - Observe $x_n$.
  - Transform it using parameters $w$ and $b$ to get $p_{w,b}(Y \mid x_n)$.
  - Sample $y_n \sim p_{w,b}(Y \mid x_n)$.

Probabilistic story with regularization:

- Use hyperparameters $\alpha$ to define a prior distribution over random variables $W$, $p_\alpha(W)$.
- Sample $w \sim p_\alpha(W)$.
- For $n \in \{1, \ldots, N\}$:
  - Observe $x_n$.
  - Transform it using parameters $w$ and $b$ to get $p_{w,b}(Y \mid x_n)$.
  - Sample $y_n \sim p_{w,b}(Y \mid x_n)$.
Maximum a Posteriori (MAP) Estimation

\[
(\hat{w}, b) = \arg \max_{w, b} \log p_\alpha(w) + \sum_{n=1}^{N} \log p_{w, b}(y_n | x_n)
\]

- \(\log \text{prior}\)
- \(\log \text{likelihood}\)
Maximum a Posteriori (MAP) Estimation

$$(\hat{\mathbf{w}}, b) = \arg\max_{\mathbf{w},b} \log p_{\alpha}(\mathbf{w}) + \sum_{n=1}^{N} \log p_{\mathbf{w},b}(y_n | \mathbf{x}_n)$$

Typical assumption is that each weight is independent of the others.

$$p_{\alpha}(\mathbf{W}) = \prod_{j} p_{\alpha}(W_j)$$
Maximum a Posteriori (MAP) Estimation

\[
(\hat{\mathbf{w}}, b) = \arg\max_{\mathbf{w}, b} \log p_\alpha(\mathbf{w}) + \sum_{n=1}^{N} \log p_{\mathbf{w}, b}(y_n | \mathbf{x}_n)
\]

Typical assumption is that each weight is independent of the others.

\[
p_\alpha(\mathbf{W}) = \prod_j p_\alpha(W_j)
\]

Option 1: let \( p_\alpha(W_j) \) be a zero-mean Gaussian distribution with standard deviation \( \alpha \).

\[
\log p_\alpha(\mathbf{w}) = -\frac{1}{2\alpha^2} \| \mathbf{w} \|_2^2 + \text{constant}
\]
Maximum a Posteriori (MAP) Estimation

\[
(\hat{w}, b) = \arg\max_{w,b} \log p_\alpha(w) + \sum_{n=1}^{N} \log p_{w,b}(y_n | x_n)
\]

Typical assumption is that each weight is independent of the others.

\[
p_\alpha(W) = \prod_j p_\alpha(W_j)
\]

Option 1: let \( p_\alpha(W_j) \) be a zero-mean Gaussian distribution with standard deviation \( \alpha \).

\[
\log p_\alpha(w) = -\frac{1}{2\alpha^2} \|w\|_2^2 + \text{constant}
\]

Option 2: let \( p_\alpha(W_j) \) be a zero-location Laplace distribution with scale \( \alpha \).

\[
\log p_\alpha(w) = -\frac{1}{\alpha} \|w\|_1 + \text{constant}
\]
Probabilistic Story: $L_2$-Regularized Logistic Regression

$$x \times w \times \sum b \times \text{logistic} \rightarrow Y$$

$$\text{MAP} \rightarrow \hat{w}, \hat{b}$$
Why Go Probabilistic?

- Interpret the classifier’s activation function as a (log) probability (density), which encodes uncertainty.
- Interpret the regularizer as a (log) probability (density), which encodes uncertainty.
- Leverage theory from statistics to get a better understanding of the guarantees we can hope for with our learning algorithms.
- Change your assumptions, turn the optimization-crank, and get a new machine learning method.

The key to success is to tell a probabilistic story that’s reasonably close to reality, including the prior(s).