Machine Learning (CSE 446): Probabilistic Generative Machine Learning

Sham M Kakade

University of Washington
cse446-staff@cs.washington.edu

Quick Review

- New view of log and squared loss functions: they are log likelihood functions!
- New view of regularized logistic/linear regression: maximize $\log p(\text{parameters}) + \log p(\text{outputs} \mid \text{inputs})$

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.

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

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

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

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

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

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

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

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

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

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

In the **generative** version, the model defines the *joint* distribution $p(X, Y)$.

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

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

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

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

In the **generative** version, the model defines the *joint* distribution $p(X, Y)$.

What we saw earlier this week was the **conditional** version.

Chain Rule of Probabilities

For any ordering of M random variables $V_1, \ldots, V_M$:

$$p(V_1, V_2, \ldots, V_M) = p(V_1) \cdot p(V_2 \mid V_1) \cdots p(V_M \mid V_1, \ldots, V_{M-1})$$
$$= \prod_{m=1}^{M} p(V_m \mid V_1, \ldots, V_{m-1})$$

Chain Rule of Probabilities

For any ordering of M random variables $V_1, \ldots, V_M$:

$$p(V_1, V_2, \ldots, V_M) = p(V_1) \cdot p(V_2 \mid V_1) \cdots p(V_M \mid V_1, \ldots, V_{M-1})$$
$$= \prod_{m=1}^{M} p(V_m \mid V_1, \ldots, V_{m-1})$$

Consider r.v.s Y (our output variable) and $X_1, \ldots, X_d$ (our d feature inputs).

$$p(Y, X_1, X_2, \ldots, X_d) = p(Y) \cdot \prod_{j=1}^{d} p(X_j \mid Y, X_1, \ldots, X_{j-1})$$

Chain Rule of Probabilities

For any ordering of M random variables $V_1, \ldots, V_M$:

$$p(V_1, V_2, \ldots, V_M) = p(V_1) \cdot p(V_2 \mid V_1) \cdots p(V_M \mid V_1, \ldots, V_{M-1})$$
$$= \prod_{m=1}^{M} p(V_m \mid V_1, \ldots, V_{m-1})$$

Consider r.v.s Y (our output variable) and $X_1, \ldots, X_d$ (our d feature inputs).

$$p(Y, X_1, X_2, \ldots, X_d) = p(Y) \cdot \prod_{j=1}^{d} p(X_j \mid Y, X_1, \ldots, X_{j-1})$$
$$\stackrel{\text{naïve assumption}}{=} p(Y) \cdot \prod_{j=1}^{d} p(X_j \mid Y)$$

We'll stick with the convention that $y \in \{-1, +1\}$ but assume that "binary feature" means values in $\{0, 1\}$.

Naïve Bayes Classification

$$f^{(\mathsf{BO})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} \mathcal{D}(\mathbf{x}, y)$$

$$f^{(\mathsf{NB})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} p(\mathbf{x}, y)$$

$$= \operatorname*{argmax}_{y \in \{-1,+1\}} p(Y = y) \cdot \prod_{j=1}^{d} p(X_j = \mathbf{x}[j] \mid Y = y)$$

Naïve Bayes Classification

$$f^{(\mathsf{BO})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} \mathcal{D}(\mathbf{x}, y)$$

$$f^{(\mathsf{NB})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} p(\mathbf{x}, y)$$

$$= \operatorname*{argmax}_{y \in \{-1,+1\}} p(Y = y) \cdot \prod_{j=1}^{d} p(X_j = \mathbf{x}[j] \mid Y = y)$$

It's called "naïve" because of the assumption that each X_j is conditionally independent of the others, given $Y = y$.

Naïve Bayes Classification

$$f^{(\mathsf{BO})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} \mathcal{D}(\mathbf{x}, y)$$

$$f^{(\mathsf{NB})}(\mathbf{x}) = \operatorname*{argmax}_{y \in \{-1,+1\}} p(\mathbf{x}, y)$$

$$= \operatorname*{argmax}_{y \in \{-1,+1\}} p(Y = y) \cdot \prod_{j=1}^{d} p(X_j = \mathbf{x}[j] \mid Y = y)$$

It's called "naïve" because of the assumption that each X_j is conditionally independent of the others, given $Y = y$.

It's called "Bayes" because we can motivate it using Bayes' rule . . .

The "Bayes" Part

It's not really about the Bayes optimal classifier, or about Bayesian probability!

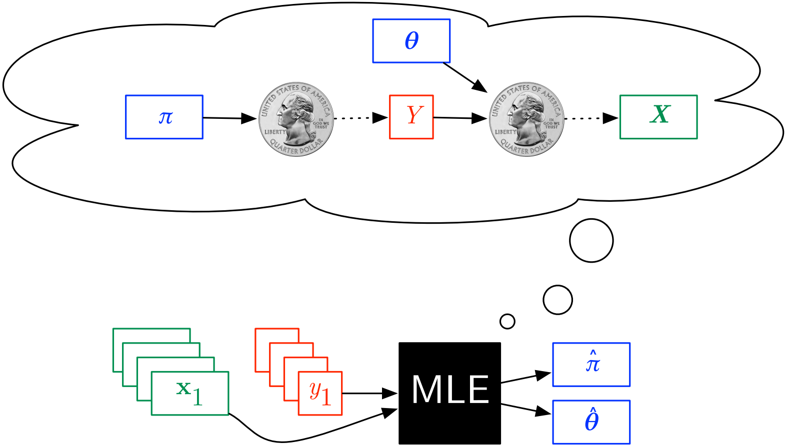
Motivation: we want $\hat{y} = \operatorname{argmax}_y p(Y = y \mid \boldsymbol{X} = \mathbf{x})$.

Bayes' rule:

$$p(Y \mid \boldsymbol{X}) = \frac{\overbrace{p(Y)}^{\text{prior}} \cdot \overbrace{p(\boldsymbol{X} \mid Y)}^{\text{likelihood}}}{\underbrace{p(\boldsymbol{X})}_{\text{evidence}}}$$

$$\begin{aligned}
\hat{y} &= \operatorname*{argmax}_y p(Y = y \mid \boldsymbol{X} = \mathbf{x}) \\
&= \operatorname*{argmax}_y \frac{p(Y = y) \cdot p(\boldsymbol{X} = \mathbf{x} \mid Y = y)}{p(\boldsymbol{X} = \mathbf{x})} \\
&= \operatorname*{argmax}_y p(Y = y) \cdot p(\boldsymbol{X} = \mathbf{x} \mid Y = y)
\end{aligned}$$

Naïve Bayes Illustrated



Naïve Bayes: Probabilistic Story (All Binary Features)

1. Sample Y according to a Bernoulli distribution where:

$$p(Y = +1) = \pi$$
$$p(Y = -1) = 1 - \pi$$

2. For each feature X_j:
   - Sample X_j according to a Bernoulli distribution where:

$$p(X_j = 1 \mid Y = y) = \theta_{X_j|y}$$
$$p(X_j = 0 \mid Y = y) = 1 - \theta_{X_j|y}$$

Naïve Bayes: Probabilistic Story (All Binary Features)

1. Sample Y according to a Bernoulli distribution where:

$$p(Y = +1) = \pi$$
$$p(Y = -1) = 1 - \pi$$

2. For each feature X_j:
   - Sample X_j according to a Bernoulli distribution where:

$$p(X_j = 1 \mid Y = y) = \theta_{X_j|y}$$
$$p(X_j = 0 \mid Y = y) = 1 - \theta_{X_j|y}$$

$1 + 2d$ parameters to estimate: $\pi, \{\theta_{X_j|+1}, \theta_{X_j|-1}\}_{j=1}^{d}$.

Naïve Bayes: Maximum Likelihood Estimation (All Binary Features)

In general, for a Bernoulli with parameter π, if the observations are $o_1, \ldots, o_N$:

$$\hat{\pi} = \frac{\text{count}(+1)}{\text{count}(+1) + \text{count}(-1)} = \frac{|\{n : o_n = +1\}|}{N}$$

Naïve Bayes: Maximum Likelihood Estimation (All Binary Features)

In general, for a Bernoulli with parameter π, if the observations are $o_1, \ldots, o_N$:

$$\hat{\pi} = \frac{\text{count}(+1)}{\text{count}(+1) + \text{count}(-1)} = \frac{|\{n : o_n = +1\}|}{N}$$

In general, for a conditional Bernoulli for $p(A \mid B)$, if the observations are $(a_1, b_1), \ldots, (a_N, b_N)$:

$$\hat{\theta}_{A|+1} = \frac{\text{count}(A = 1, B = +1)}{\text{count}(B = +1)} = \frac{|\{n : a_n = 1 \wedge b_n = +1\}|}{|\{n : b_n = +1\}|}$$

$$\hat{\theta}_{A|-1} = \frac{\text{count}(A = 1, B = -1)}{\text{count}(B = -1)} = \frac{|\{n : a_n = 1 \wedge b_n = -1\}|}{|\{n : b_n = -1\}|}$$

Naïve Bayes: Maximum Likelihood Estimation (All Binary Features)

In general, for a Bernoulli with parameter π, if the observations are $o_1, \ldots, o_N$:

$$\hat{\pi} = \frac{\text{count}(+1)}{\text{count}(+1) + \text{count}(-1)} = \frac{|\{n : o_n = +1\}|}{N}$$

In general, for a conditional Bernoulli for $p(A \mid B)$, if the observations are $(a_1, b_1), \ldots, (a_N, b_N)$:

$$\hat{\theta}_{A|+1} = \frac{\text{count}(A = 1, B = +1)}{\text{count}(B = +1)} = \frac{|\{n : a_n = 1 \wedge b_n = +1\}|}{|\{n : b_n = +1\}|}$$

$$\hat{\theta}_{A|-1} = \frac{\text{count}(A = 1, B = -1)}{\text{count}(B = -1)} = \frac{|\{n : a_n = 1 \wedge b_n = -1\}|}{|\{n : b_n = -1\}|}$$

So for naïve Bayes' parameters:

- $\hat{\pi} = \dfrac{|\{n : y_n = +1\}|}{N}$

Naïve Bayes: Maximum Likelihood Estimation (All Binary Features)

In general, for a Bernoulli with parameter π, if the observations are $o_1, \ldots, o_N$:

$$\hat{\pi} = \frac{\text{count}(+1)}{\text{count}(+1) + \text{count}(-1)} = \frac{|\{n : o_n = +1\}|}{N}$$

In general, for a conditional Bernoulli for $p(A \mid B)$, if the observations are $(a_1, b_1), \ldots, (a_N, b_N)$:

$$\hat{\theta}_{A|+1} = \frac{\text{count}(A = 1, B = +1)}{\text{count}(B = +1)} = \frac{|\{n : a_n = 1 \wedge b_n = +1\}|}{|\{n : b_n = +1\}|}$$

$$\hat{\theta}_{A|-1} = \frac{\text{count}(A = 1, B = -1)}{\text{count}(B = -1)} = \frac{|\{n : a_n = 1 \wedge b_n = -1\}|}{|\{n : b_n = -1\}|}$$

So for naïve Bayes' parameters:

- $\hat{\pi} = \frac{|\{n : y_n = +1\}|}{N}$
- For each j and each $y \in \{-1, +1\}$: $\hat{\theta}_{j,y} = \frac{|\{n : y_n = y \wedge \mathbf{x}_n[j] = 1\}|}{|\{n : y_n = y\}|}$

Beyond Binary Features

For X_j that are not binary, there are many options for $p(X_j \mid Y = +1)$ and $p(X_j \mid Y = -1)$.

Some often-used ones are:

- For continuous X_j, define two Gaussian densities with parameters $\langle \mu_{X_j|+1}, \sigma^2_{X_j|+1} \rangle$ and $\langle \mu_{X_j|-1}, \sigma^2_{X_j|-1} \rangle$.
- For nonnegative integer X_j, define two Poisson distributions with parameters $\lambda_{X_j|+1}$ and $\lambda_{X_j|-1}$.