Generative Models

Peng Lingwei

August 22, 2019

Contents

24 Generative Models

1. Distribution free learning framework;
2. Generative approach: parametric density estimation;
3. When solving a given problem, try to avoid a more general problem as an intermediate step.

24.1 MAXIMUM LIKELIHOOD ESTIMATOR

For a 0–1 distribution, and the true parameter is θ^*. We estimate $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} x_i$, then

$$\mathbb{P}\left\{\left|\hat{\theta} - \theta^*\right| \leq \sqrt{\frac{\log(1/\delta)}{2m}}\right\} \geq 1 - \delta$$

Maximum likelihood estimation function:

$$L(S;\theta) = \log\left(\prod p_\theta(x_i)\right) = \sum_{i=1}^{m} \log(p_\theta(x_i))$$

The maximum likelihood estimator uses the loss $l(\theta, x) = -\log(p_\theta(x))$ and estimates θ by ERM rules

$$\arg\min_\theta \sum_{i=1}^{m}(-\log(p_\theta(x_i))) = \arg\max_\theta \sum_{i=1}^{m} \log(p_\theta(x_i))$$

The true risk of a parameter θ becomes (Realizable cases, the true distribution is in the assumption distribution class):

$$\begin{aligned}
&\mathbb{E}_x\left[l(\theta, x)\right] = -\sum_x p_{\theta^*}(x) \log p_\theta(x) \\
=&\sum_x p_{\theta^*}(x) \log\left(\frac{p_{\theta^*(x)}}{p_\theta(x)}\right) + \sum_x p_{\theta^*}(x) \log\left(\frac{1}{p_{\theta^*}(x)}\right) \\
=&D_{RE}\left[p_{\theta^*} \| p_\theta\right] + H(p_{\theta^*})
\end{aligned}$$

D_{RE} is called the relative entropy, and H is called the entropy function.

$$D_{RE}(p\|q) = \mathbb{E}_p\left[\log\frac{p}{q}\right] \geq -\log \mathbb{E}_p\left[\frac{q}{p}\right] = -\log q \geq 0$$

In Gaussian variable of unit variance,

$$\begin{aligned}
&\mathbb{E}_{x\sim N(\mu^*,1)}\left[l(\hat{\mu}, x) - l(\mu^*, x)\right] = \mathbb{E}_{x\sim N(\mu^*,1)} \log\left(\frac{p_{\mu^*}(x)}{p_{\hat{\mu}}(x)}\right) \\
=&\mathbb{E}_{x\sim N(\mu^*,1)}\left(-\frac{1}{2}(x-\mu^*)^2 + \frac{1}{2}(x-\hat{\mu})^2\right) \\
=&\frac{1}{2}\left(r\hat{\mu}^2 - \mu^{*2} + 2(\mu^* - \hat{\mu})\mathbb{E}_{x\sim N(\mu^*,1)}(x)\right) \\
=&\frac{1}{2}\left(\hat{\mu}^2 - \mu^{*2} + 2(\mu^* - \hat{\mu})\mu^*\right) = \frac{1}{2}(\hat{\mu} - \mu^*)^2.
\end{aligned}$$

$$\mathbb{P}\left\{|\mu-\mu^*|\leq\sqrt{\frac{\log(1/\delta)}{2m}}\right\}\geq 1-\delta\Rightarrow\mathbb{P}\left\{\frac{1}{2}(\hat{\mu}-\mu^*)^2\leq\frac{\log(1/\delta)}{4m}\right\}\geq 1-\delta$$

In some situations, the maximum likelihood estimator clearly overfits. Consider a Bernoulli random variable X and let $P(X=1)=\theta^*$. We can guarantee $|\theta-\theta^*|$ is small with high probability.But we can show that the true log-loss may be large.

$$\mathbb{P}\left(\forall x\in S, x=0|\theta^*\right)=(1-\theta^*)^m\geq e^{-2\theta^* m}(\geq 0.5\ if\ m\leq\frac{\ln 2}{2\theta^*})$$

In this situation, the maximum likelihood rule will set $\hat{\theta}=0$, and the true error is

$$\mathbb{E}_{x\sim\theta^x}\left[l(\hat{\theta},x)\right]=\theta^* l(\hat{\theta},1)+(1-\theta^*)l(\hat{\theta},0)=\theta^*\log(1/\hat{\theta})+(1-\theta^*)\log(1/(1-\hat{\theta}))=\infty$$

We can use regularization for maximum likelihood to avoid this problem:

$$L_S(\theta)=\frac{1}{m}\sum_{i=1}^m\log(1/p_\theta(x_i))+\frac{1}{m}(\log(1/\theta)+\log(1/(1-\theta)))$$

1. $\hat{\theta}=\frac{1}{m+2}\left(1+\sum_{i=1}^m x_i\right).$
2. $$\left|\hat{\theta}-\theta^*\right|\leq\left|\hat{\theta}-\mathbb{E}(\hat{\theta})\right|+\left|\mathbb{E}(\hat{\theta})-\theta^*\right|=\left|\hat{\theta}-\frac{1+m\theta^*}{m+2}\right|+\left|\frac{1-2\theta^*}{m+2}\right|$$
$$=\frac{m}{m+2}\left|\frac{1}{m}\sum_{i=1}^m x_i-\theta^*\right|+\left|\frac{1-2\theta^*}{m+2}\right|\leq\frac{m}{m+2}\left|\frac{1}{m}\sum_{i=1}^m x_i-\theta^*\right|+\frac{1}{m+2}$$
$$\mathbb{P}\left\{\left|\hat{\theta}-\theta^*\right|\leq\frac{m}{m+2}\sqrt{\frac{\log(1/\delta)}{2m}}+\frac{1}{m+2}\right\}\geq 1-\delta$$
3. $$\begin{aligned}&\mathbb{E}_x\left[l(\theta,x)\right]=-\theta^*\ln(\theta)-(1-\theta^*)\ln(1-\theta)\\ &\leq\max\left\{-\ln(\theta),-\ln(1-\theta)\right\}\leq\ln(m+2)\end{aligned}$$

24.2 NAIVE BAYES

Consider the problem of predicting a label $y\in\{0,1\}$ on the basis of a vector of features $\vec{x}=(x_1,\ldots,x_d)\in\{0,1\}^d$. Then the bayes optimal classifier is

$$h_{Bayes}(\vec{x})=\arg\max_{y\in\{0,1\}}P\left[Y=y|X=\vec{x}\right].$$

$\forall \vec{x} \in \{0,1\}^d$, we need calculate 2^d parameters $P\left[Y=1|X=\vec{x}\right]$. We can use Naive Bayes approach to simplify

$$\begin{aligned} h_{Bayes}(\vec{x}) &= \arg\max_{y\in\{0,1\}} P\left[Y=y|X=\vec{x}\right] \\ &= \arg\max_{y\in\{0,1\}} P\left[Y=y\right] P\left[X=\vec{x}|Y=y\right]/P\left[X=\vec{x}\right] \\ &= \arg\max_{y\in\{0,1\}} P[Y=y]\prod_{i=1}^{d} P\left[X_i=x_i|Y=y\right] \end{aligned}$$

Then, we only need estimate 2d+1 parameters.

24.3 LINEAR DISCRIMINANT ANALYSIS

Let $P[Y=1]=p, P[Y=0]=1-p$. And assume that the conditional probability of X given Y is a Gaussian distribution.Then, $h_{Bayes}(\vec{x}) =$ iff

$$\log\left(\frac{P[Y=1]P[X=\vec{x}|Y=1]}{P[Y=0]P[X=\vec{x}|Y=0]}\right) > 0$$

$$\frac{\mu}{2}(\vec{x}-\vec{\mu}_0)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_0) - \frac{1-\mu}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_1) > 0$$

If $\mu = 0.5$, the bound is a linear and we call it linear discriminant.

24.4 LATENT VARIABLES AND THE EM ALGORITHM

We construct a instance space $\mathcal{X}$ with latent random variables $\mathcal{Y} = \{1,\ldots,k\}$, and $P\left[Y=y\right]=c_y$. Second, we choose $\vec{x}$ on the basis of the value of Y according to a Gaussian distribution

$$P\left[X=\vec{x}|Y=y\right] = \frac{1}{(2\pi)^{d/2}\left|\Sigma_y\right|^{1/2}}\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_y)^T\Sigma_y^{-1}(\vec{x}-\vec{\mu}_y)\right).$$

Then X is a mixed Gaussian distribution

$$P\left[X=\vec{x}\right] = \sum_{y=1}^{k} P\left[Y=y\right]P\left[X=\vec{x}|Y=y\right]$$

The parameters are $c_y, \vec{\mu}_y, \Sigma_y$, where $y=1,\ldots,k$. The maximum-likelihood estimator is therefore the solution of the maximization problem

$$\arg\max_{c_y,\vec{\mu}_y,\Sigma_y} \sum_{i=1}^{m}\log\left(\sum_{y=1}^{k} P_{c_y,\vec{\mu}_y,\Sigma_y}\left[X=\vec{x}_i, Y=y\right]\right)$$

Now we put aside the mixed Gaussian distribution. Define $Q_{i,y} = P\left[Y=y|\vec{x}_i\right]$, then

$$F(Q,\vec{\theta}) = \sum_{i=1}^{m}\sum_{y=1}^{k} Q_{i,y}\log(P_{\vec{\theta}}\left[X=\vec{x}_i, Y=y\right]).$$

Definition 1. *(EM)*

1. *Expectation Step:* $Q_{i,y}^{(t+1)} = P_{\vec{\theta}^{(t)}}\left[Y = y | X = \vec{x}_i\right]$;

2. *Maximization Step:* $\vec{\theta}^{(t+1)} = \arg\max_{\vec{\theta}} F(Q^{(t+1)}, \vec{\theta})$.

Let $G(Q, \theta) = F(Q, \theta) - \sum_{i=1}^m \sum_{y=1}^k Q_{i,y} \log(Q_{i,y})$

Lemma 1. *The EM procedure can be rewritten as*

$$Q^{(t+1)} = \arg\max_Q G(Q, \vec{\theta}^{(t)})$$

$$\vec{\theta}^{(t+1)} = \arg\max_\theta G(Q^{(t+1)}, \vec{\theta})$$

Furthermore, $G(Q^{(t+1)}, \vec{\theta}^{(t)}) = L(\vec{\theta}^{(t)})$.

Proof. First we have $\arg\max_{\vec{\theta}} G(Q^{(t+1)}, \theta) = \arg\max_{\vec{\theta}} F(Q^{(t+1)}, \vec{\theta})$.

$$\begin{aligned} G(Q, \vec{\theta}) &= \sum_{i=1}^m \sum_{y=1}^k Q_{i,y} \log\left(\frac{P_{\vec{\theta}}\left[X = \vec{x}_i, Y = y\right]}{Q_{i,y}}\right) \\ &\leq \sum_{i=1}^m \log\left(\sum_{y=1}^k Q_{i,y} \frac{P_{\vec{\theta}}\left[X = \vec{x}_i, Y = y\right]}{Q_{i,y}}\right) \\ &= \sum_{i=1}^m \log\left(P_{\vec{\theta}}\left[X = \vec{x}_i\right]\right) = L(\vec{\theta}) \end{aligned}$$

If $Q_{i,y} = P_{\vec{\theta}}\left[Y = y | X = \vec{x}_i\right]$, it's easy to verify that $G(Q, \vec{\theta}) = L(\vec{\theta})$. □

Theorem 1. $L(\theta^{(t+1)}) \geq L(\theta^{(t)})$.

Proof. $L(\vec{\theta}^{(t+1)}) = G(Q^{(t+2)}, \vec{\theta}^{(t+1)}) \geq G(Q^{(t+1)}, \vec{\theta}^{(t+1)}) \geq G(Q^{(t+1)}, \vec{\theta}^{(t)}) = L(\theta^{(t)})$ □

Then we go back to mixed Gaussian distribution. We assume that $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = I$.

1. Expectation step:

$$P_{\theta^{(t)}}\left[Y = y | X = \vec{x}_i\right] = \frac{1}{Z_i} P_{\theta^{(t)}}[Y = y] P_{\theta^{(t)}}[X = \vec{x}_i | Y = y] = \frac{1}{Z_i} c_y^{(t)} \exp\left(\frac{1}{2}\|\vec{x}_i - \vec{\mu}_y^{(t)}\|^2\right).$$

2. Maximumization step:

$$\sum_{i=1}^m \sum_{y=1}^k P_{\theta^{(\vec{t})}}\left[Y = y | X = \vec{x}_i\right] \left(\log(c_y) - \frac{1}{2}\|\vec{x}_i - \vec{\mu}_y\|^2\right)$$

$$\vec{\mu}_y = \sum_{i=1}^m P_{\theta^{(\vec{t})}}\left[Y = y | X = \vec{x}_i\right] \vec{x}_i$$

$$c_y = \frac{\sum_{i=1}^m P_{\vec{\theta}^{(t)}}\left[Y = y | X = \vec{x}_i\right]}{\sum_{y'=1}^k \sum_{i=1}^m P_{\vec{\theta}^{(t)}}\left[Y = y' | X = \vec{x}_i\right]}$$

24.5 BAYESIAN REASONING

1. Maximum likelihood estimator assumes that parameter θ is fixed but unknow;
2. Bayesian approach: θ is a random variable, $P[\theta]$ is called prior distribution.

$$P\left[X=x\right]=\sum_{\theta}P\left[X=x,\theta\right]=\sum_{\theta}P[\theta]P[X=x|\theta]$$

or

$$P[X=x]=\int_{\theta}P[\theta]P[X=x|\theta]d\theta.$$

In the Bayesian framework, X and S are note independent anymore.

$$P[\theta|S]=\frac{P[S|\theta]P[\theta]}{P[S]}=\frac{1}{P[S]}\prod_{i=1}^{m}P[X=x_i|\theta]P[\theta]$$

$$\begin{aligned}P\left[X=x|S\right]=&\sum_{\theta}P[X=x|\theta,S]P[\theta|S]=\sum_{\theta}P[X=x|\theta]P[\theta|S]\\=&\frac{1}{P[\theta]}\sum_{\theta}P[X=x|\theta]\prod_{i=1}^{m}P[X=x_i|\theta]P[\theta]\end{aligned}$$

In binary classification problem, if θ is uniform, we have

$$P\left[X=1|S\right]\propto\int\theta^{1+\sum_i x_i}(1-\theta)^{\sum_{i=1}(1-x_i)}d\theta$$

$$\int\theta^{A}(1-\theta)^{B}d\theta=\frac{B}{A+1}\int\theta^{A+1}(1-\theta)^{B-1}d\theta=\cdots=\frac{A!B!}{(A+B)!}\int\theta^{A+B}d\theta$$

$$\frac{P\left[X=1|S\right]}{P\left[X=0|S\right]}=\frac{(1+\sum_{i=1}^{m}x_i)!(\sum_{i=1}^{m}(1-x_i))!}{(\sum_{i=1}^{m}x_i)!(1+\sum_{i=1}^{m}(1-x_i))!}=\frac{1+\sum_{i=1}^{m}x_i}{1+\sum_{i=1}^{m}(1-x_i)}$$

$$\Rightarrow P\left[X=1|S\right]=\frac{1+\sum_{i=1}^{m}x_i}{m+2}$$

Bayesian prediction adds “pseudoexamples” to the training set.