Lecture-15: Generative models

1 Introduction

We followed a *discriminative* approach in which our goal is not to learn the underlying distribution but rather to learn an accurate predictor. That is, a sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ is *i.i.d.* with a fixed and unknown distribution $\mathcal{D}$.

An alternative approach where we assume that the underlying distribution over the data has a specific parametric form and our goal is to estimate the parameters of the model is called the *generative* approach, and this task is called *parametric density estimation*. That is, we assume that the distribution $\mathcal{D}$ has a density from a parametric family $(p_\theta : \theta \in \Theta)$ with unknown parameter $\theta \in \Theta$.

The discriminative approach has the advantage of directly optimizing the prediction accuracy instead of learning the underlying distribution.

Principle 1 (Vladimir Vapnik). Principle for solving problems using a restricted amount of information: *When solving a given problem, try to avoid a more general problem as an intermediate step.*

Remark 1. If we succeed in learning the underlying distribution accurately, we can predict labels for new examples using the Bayes optimal classifier. The problem is that it is usually more difficult to learn the underlying distribution than to learn an accurate predictor. However, in some situations, it is reasonable to adopt the generative learning approach. For example, sometimes it is computationally easier to estimate the parameters of the model than to learn a discriminative predictor. Additionally, in some cases we do not have a specific task at hand but rather would like to model the data either for making predictions at a later time without having to retrain a predictor or for the sake of interpretability of the data.

If we knew the true parameter θ, then the distribution p_θ is known and a Bayes or ML estimator can be used. For unknown parameter, we estimate $\hat{\theta}_m$ from the sample S. For a Bayesian estimate, the distribution $p_{\hat{\theta}_m}$ can be used as the true distribution for a new observation to estimate the label y. For maximum likelihood estimate, one selects the parameter θ that maximizes the likelihood of the observations. That is,

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{m} \log p_\theta(x_i, y_i).$$

The new observation is assumed to be from the density $p_{\hat{\theta}_{\mathrm{ML}}}$.

Principle 2 (Vladimir Vapnik). As the sample sizes grows large, each of the descriptive and generative approaches reach their asymptotic value of the generalized risk.

Remark 2. Each approach may have a different asymptotic generalized risk. Asymptotic performance of discriminative approaches is generally better than those of generative approaches. However, the asymptotic performance is reached slower for the discriminative approaches when compared to generative approaches. The asymptotic performance is reached in $O(m)$ steps for discriminative approaches, as compared to $O(\ln m)$ steps for generative approaches. Therefore, for small sample size, generative approaches may outperform discriminative approaches, which get better for large sample size.

2 Maximum likelihood estimator

Definition 2.1. If a sample $S \in \mathcal{X}^m$ is generated *i.i.d.* from a density p_θ, then the log-likelihood of sample S is defined as $L(S;\theta) \triangleq \ln \prod_{i=1}^{m} p_\theta(x_i)$. The maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}} \triangleq \arg\max_\theta L(S;\theta)$.

Example 2.2. Consider an *i.i.d.* Bernoulli sample $S \in \mathcal{X}^m$ where $\mathcal{X} = \{0,1\}$ and the underlying probability of $P\{X_1 = 1\} = \theta$ for some $\theta \in \Theta = [0,1]$. We are interested in estimating θ from the sample S. We can write the probability of observation S as

$$P_\theta(S) = \prod_{i=1}^{m} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i\in[m]} x_i}(1-\theta)^{\sum_{i\in[m]}(1-x_i)}.$$

The log-likelihood of sample S for a parameter θ is

$$L(S;\theta) \triangleq \ln P_\theta(S) = \ln\theta \sum_{i\in[m]} x_i + \ln\bar{\theta} \sum_{i\in[m]} \bar{x}_i.$$

The maximum likelihood estimator of S is $\hat{\theta} \in \arg\max L(S;\theta)$

2.1 Empirical risk minimization

Definition 2.3. Given a parameter $\theta \in \Theta$ and observation $x \in \mathcal{X}$ generated *i.i.d.* from density p_θ, we define a loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}_+$ as the negative of log-likelihood of observation. That is, $\ell(\theta,x) \triangleq -\log_2 p_\theta(x)$.

Remark 3. Maximum likelihood estimator is minimizing the empirical risk with respect to loss function ℓ defined in Definition 2.3.

Remark 4. If the observation $x \in \mathcal{X}$ is *i.i.d.* with a true density p, then the generalized risk of parameter θ is

$$\mathbb{E}\ell(\theta,x) = -\int_{x\in\mathcal{X}} p(x)\log_2 p_\theta(x) = D(p\|p_\theta) + H(p).$$

For discrete space $\mathcal{X}$, the relative entropy $D(p\|p_\theta) \geqslant 0$ and equal to zero when $p_\theta = p$.

Remark 5. If the true distribution $\mathcal{D} = p_{\theta_0}$ for some $\theta_0 \in \Theta$, then $D(p\|p_\theta) = D(p_{\theta_0}\|p_\theta)$, and this loss function is minimized for $\theta = \theta_0$. It shows that if the underlying distribution indeed has a parametric form, then by choosing the correct parameter we can make the risk be the entropy of the distribution.

Remark 6. This expression underscores how our generative assumption affects our density estimation, even in the limit of infinite data. If the underlying distribution is not of the assumed parametric form, even the best parameter leads to an inferior model and the sub-optimality is measured by the relative entropy divergence.

3 EM algorithm

Assumption 3.1. Sample $S \in \mathcal{X}^m$ is generated from a specific parametric distribution, generated by latent (hidden) random variables over discrete state space $\mathcal{Y}$. Specifically, let $\theta \in \Theta$ be the parameters for the joint distribution over state space $\mathcal{X} \times \mathcal{Y}$.

Example 3.2 (Mixture of Gaussian random variables). Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = [k]$ where the probability mass function of random variable Y is $c \in \mathcal{M}(\mathcal{Y})$ and the conditional density of $x \in \mathcal{X}$ for each latent variable $y \in \mathcal{Y}$ is

$$p_y(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma_y}} \exp\Big(-\frac{1}{2}(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)\Big),$$

for mean vector $\mu_y \in \mathcal{X}$ and covariance matrix $\Sigma_y \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$. It follows that density of observation $x \in \mathcal{X}$ is $f_X(x) = \sum_{y\in\mathcal{Y}} c_y p_y(x)$. Here, the parameters of joint distribution are (c_y, μ_y, Σ_y) for each $y \in \mathcal{Y}$.

Definition 3.3. The log-likelihood of an observation is defined as $\log p_\theta(x) = \log\Big(\sum_{y\in\mathcal{Y}} p_\theta(x,y)\Big)$. Given an *i.i.d.* sample $S \in \mathcal{X}^m$, the log-likelihood of sample is defined as

$$L(\theta) \triangleq \log\prod_{i=1}^{m} p_\theta(x_i) = \sum_{i=1}^{m}\log\Big(\sum_{y\in\mathcal{Y}} p_\theta(x_i,y)\Big).$$

Problem 1 (ML estimator). The maximum-likelihood estimator is the solution of the maximization problem

$$\hat{\theta}_{\mathrm{ML}} \triangleq \arg\max_{\theta} L(\theta) = \arg\max_{\theta}\sum_{i=1}^{m}\log\Big(\sum_{y\in\mathcal{Y}} p_\theta(x_i,y)\Big).$$

Remark 7. In many situations, the summation inside the log makes the ML optimization problem computationally hard.

Remark 8. Using the log-sum inequality which gives $a\log\frac{a}{b} \leqslant \sum_i a_i \log\frac{a_i}{b_i}$, we can write

$$\log\Big(\sum_{y\in\mathcal{Y}} p_\theta(x_i,y)\Big) = \log\Big(\sum_{y\in\mathcal{Y}} P_\theta(\{Y=y\}\mid\{X=x_i\})p_\theta(x_i)\Big) \leqslant$$

Remark 9. The Expectation-Maximization (EM) algorithm, due to Dempster, Laird, and Rubin, is an iterative procedure for searching a local maximum of log-likelihood $L(\theta)$. While EM is not guaranteed to find the global maximum, it often works reasonably well in practice.

Remark 10. EM is designed for those cases in which, had we known the values of the latent variables Y , then the maximum likelihood optimization problem would have been tractable.

Definition 3.4. Consider a matrix $Q \in [0,1]^{[m]\times\mathcal{Y}}$ such that $Q_i \in \mathcal{M}(\mathcal{Y})$ is the conditional distribution for example $x_i \in \mathcal{X}$ and $i \in [m]$. For a parameter $\theta \in \Theta$, we define expected log-likelihood $F(Q,\theta)$ of a sample $S \in \mathcal{X}^m$ where ith example has distribution Q_i, written as

$$F(Q,\theta) \triangleq \mathbb{E}_Q \log p_\theta(x_i,y) = \sum_{i=1}^{m}\sum_{y\in\mathcal{Y}} Q_{i,y}\log p_\theta(x_i,y).$$

Assumption 3.5. For any matrix $Q \in [0,1]^{[m]\times\mathcal{Y}}$ the optimization problem $\arg\max_\theta F(Q,\theta)$ is tractable.

Definition 3.6 (EM algorithm). EM algorithm finds a sequence of solutions $((Q^{(1)},\theta^{(1)}),(Q^{(2)},\theta^{(2)}),\dots)$. The initial values of $Q^{(1)}$ and $\theta^{(1)}$ are usually chosen at random and the procedure terminates after the improvement in the likelihood value stops being significant. At iteration t, one constructs $(Q^{(t+1)},\theta^{(t+1)})$ by performing the following two steps.

Expectation step: $Q_{i,y}^{(t+1)} \triangleq P_{\theta^{(t)}}(\{Y=y\}\mid\{X=x_i\})$. This step is called the Expectation step, because it yields a new probability over the latent variables, which defines a new expected log-likelihood function over θ.

Maximization step: $\theta^{(t+1)} \triangleq \arg\max_\theta F(Q^{(t+1)},\theta)$. By Assumption 3.5, it is possible to efficiently find the maximizer of the expected log-likelihood, where the expectation is according to $Q^{(t+1)}$.

3.1 EM as an alternate maximization algorithm

Definition 3.7. Consider the set of distributions $\mathcal{Q} \triangleq \Big\{Q \in [0,1]^{[m]\times\mathcal{Y}} : Q_i \in \mathcal{M}(\mathcal{Y})\Big\}$, and the objective function $G : \mathcal{Q}\times\Theta \to \mathbb{R}$ defined as

$$G(Q,\theta) \triangleq F(Q,\theta) - \sum_{i=1}^{m}\sum_{y\in\mathcal{Y}} Q_{i,y}\log Q_{i,y} = F(Q,\theta) + \sum_{i=1}^{m} H(Q_i).$$

Lemma 3.8. *The EM procedure can be written as*

$$Q^{(t+1)} = \arg\max_{Q\in\mathcal{Q}} G(Q,\theta^{(t)}), \qquad \theta^{(t+1)} = \arg\max_{\theta} G(Q^{(t+1)},\theta).$$

In addition, we have $G(Q^{(t+1)},\theta^{(t)}) = L(\theta^{(t)})$.

Proof. It is clear that $\arg\max_\theta G(Q,\theta) = \arg\max_\theta F(Q,\theta)$ since $\sum_{i=1}^m H(Q_i)$ doesn't depend on parameter θ. In addition, we observe that

$$\begin{aligned} G(Q,\theta) &= \sum_{i=1}^m \Big(\sum_{y\in\mathcal{Y}} Q_{i,y} \log \frac{P_\theta(\{Y=y\} \mid \{X=x_i\})}{Q_{i,y}} + \sum_{y\in\mathcal{Y}} Q_{i,y} \log p_\theta(x_i) \Big) \\ &= -\sum_{i=1}^m D(Q_i \| P_\theta(Y \mid \{X=x_i\})) + L(\theta) \leqslant L(\theta). \end{aligned}$$

The result follows since inequality is achieved by equality for $Q_{i,y} = P_\theta(\{Y=y\} \mid \{X=x_i\})$ for all $y \in \mathcal{Y}$ and $i \in [m]$. □

Remark 11. For a fixed θ, we have $\max_Q G(Q,\theta) = L(\theta)$ and the maximizing distribution Q_i for each $i \in [m]$ is the conditional distribution of latent variable y given observation x_i.

Remark 12. The intuitive idea of EM is that we want to maximize $G(Q,\theta)$ over both Q and θ. For a known Q, the optimization problem of finding the best parameter θ is tractable when Assumption 3.5 holds. For known parameter θ, one can set $Q_{iy} = P(\{Y=y\} \mid \{X=x_i\})$. The EM algorithm therefore alternates between finding optimal parameter θ given some Q and finding optimal Q given some θ.

Corollary 3.9. *The EM procedures never decreases the log-likelihood. That is, $L(\theta^{(t+1)}) \geqslant L(\theta^{(t)})$ for all $t \in \mathbb{N}$.*

Proof. From the Lemma 3.8, we have

$$L(\theta^{(t+1)}) = G(Q^{(t+2)},\theta^{(t+1)}) \geqslant G(Q^{(t+1)},\theta^{(t+1)}) \geqslant G(Q^{(t+1)},\theta^{(t)}) = L(\theta^{(t)})$$

□

Example 3.10 (Mixture of Gaussian random variables). In Example 3.2, we assume $\Sigma_y = \mathbf{I}$ for all $y \in \mathcal{Y}$ for simplicity. Thus, we have $\theta = ((c_y, \mu_y) : y \in \mathcal{Y})$.

Expectation step: For a partition function Z_i, we can write the conditional probability as

$$Q_{iy}^{(t+1)} = P_{\theta^{(t)}}(\{Y=y\} \mid \{X=x_i\}) = \frac{1}{Z_i} P_{\theta^{(t)}}\{Y=y\} f_{X|\{Y=y\})(x_i}) = \frac{1}{Z_i} c_y^{(t)} \exp\Big(-\frac{1}{2}\left\|x_i - \mu_y^{(t)}\right\|^2\Big).$$

Maximization step: The parameter $\theta^{(t+1)}$ maximizes the step

$$F(Q^{(t+1)},\theta) = \sum_{i=1}^m \sum_{y\in\mathcal{Y}} P_{\theta^{(t)}}(\{Y=y\} \mid \{X=x_i\}) \Big(\ln c_y - \frac{1}{2}\left\|x_i - \mu_y\right\|^2 \Big).$$

Taking derivative with respect to μ_y and equating it to zero, we get $\mu_y^{(t+1)} = \sum_{i=1}^m P_{\theta^{(t)}}(\{Y=y\} \mid \{X=x_i\}) x_i$. That is, $\mu_y^{(t+1)}$ is a weighted average of the examples x_i where the weights are according to the probabilities calculated in the Expectation step. To find the optimal c we need to be more careful since we must ensure that c is a probability vector. We show that

$$c_y^{(t+1)} = \frac{\sum_{i=1}^m P_{\theta^{(t)}}(\{Y=y\} \mid \{X=x_i\})}{\sum_{z\in\mathcal{Y}} \sum_{i=1}^m P_{\theta^{(t)}}(\{Y=z\} \mid \{X=x_i\})}.$$