Generative modeling

Sep 2022

Murphy chap3

Generative or discriminative

$$p(y, \boldsymbol{x}) = P(y|\boldsymbol{x})p(\boldsymbol{x}).$$

In discriminative learning, only the first of the two terms in the product is considered; a functional form is adopted and parameterized appropriately as $P(y|\boldsymbol{x}; \boldsymbol{\theta})$. Parameters are then estimated by optimizing a cost function. The distribution of the input data is ignored. Such an approach has the advantage that simpler models can be used, especially if the input data are described by a joint probability density function (PDF) of a complex form. The disadvantage is that the input data distribution is ignored, although it can carry important information, which could be exploited to the benefit of the overall performance.

In contrast, the alternative path, known as *generative learning*, exploits the input data distribution Once more, employing the product rule, we have

$$p(y, \boldsymbol{x}) = p(\boldsymbol{x}|y)P(y),$$

Generative: models joint probability
Discriminative: models class-conditional probability P(class|data)

Generative or discriminative

$$p(y, \boldsymbol{x}) = P(y|\boldsymbol{x})p(\boldsymbol{x}).$$

In discriminative learning, only the first of the two terms in the product is considered; a functional form is adopted and parameterized appropriately as $P(y|\boldsymbol{x}; \boldsymbol{\theta})$. Parameters are then estimated by optimizing a cost function. The distribution of the input data is ignored. Such an approach has the advantage that simpler models can be used, especially if the input data are described by a joint probability density function (PDF) of a complex form. The disadvantage is that the input data distribution is ignored, although it can carry important information, which could be exploited to the benefit of the overall performance.

In contrast, the alternative path, known as *generative learning*, exploits the input data distribution Once more, employing the product rule, we have

$$p(y, \boldsymbol{x}) = p(\boldsymbol{x}|y)P(y),$$

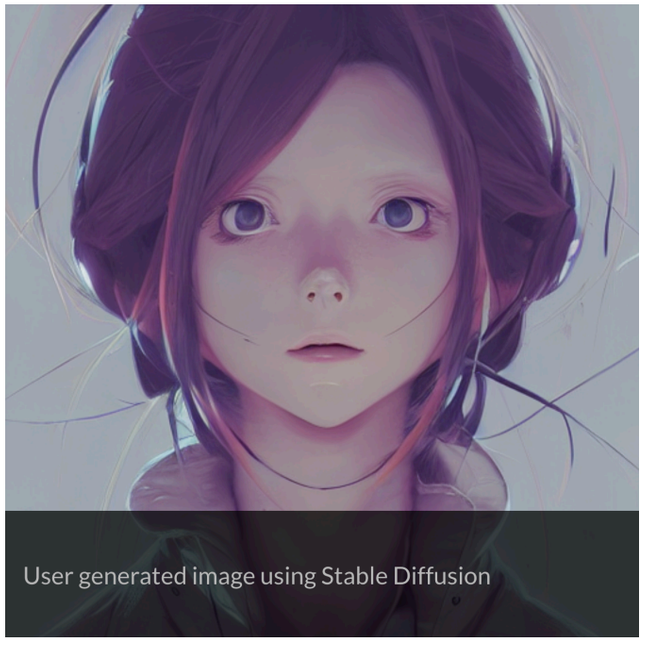
Generative: models joint probability
Discriminative: models class-conditional probability P(class|data)

Pros and cons? Murphy chap 8.6

Generative or discriminative

Generative models: e.g.

- Linear discriminative analysis (LDA)
- Naive Bayesian classifier



Discriminative models: e.g.

- Logistic regression
- Perception (binary or multi-class classifier)
- Fisher's Linear Discriminant Analysis
- Support vector machine

MAP

X: observation variable

Y: class variable $\in \{1, ...K\}$

class conditional probability $P(X = x \mid Y = k)$

Prior $P(Y = k)$

Posterior given by Bayesian formula:

$$P(Y = k \mid X) = \frac{P(X \mid Y = k)P(Y = k)}{P(X)}$$

decision rule (maximum a posterior, abb. MAP):

$$\operatorname*{argmax}_k P(Y = k \mid X) = \operatorname*{argmax}_k P(X \mid Y = k)P(Y = k)$$

$P(X)$the marginal likelihood, is independent of k

MAP

For pedagogical purposes, we will consider a very simple example of concept learning called the **number game**, based on part of Josh Tenenbaum's PhD thesis (Tenenbaum 1999). The game proceeds as follows. I choose some simple arithmetical concept C, such as "prime number" or "a number between 1 and 10". I then give you a series of randomly chosen positive examples $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn from C, and ask you whether some new test case $\tilde{x}$ belongs to C, i.e., I ask you to classify $\tilde{x}$.

$$\tilde{x} \in C?$$

Suppose, for simplicity, that all numbers are integers between 1 and 100. Now suppose I tell you "16" is a positive example of the concept. What other numbers do you think are positive?

Predictions quite vague, need more data!

$$\mathcal{D} = \{16, 8, 2, 64\}$$

Power of two OR even ?

MAP

Subjective

$$h_{\text{two}} = \{2, 4, 8, 16, 32, 64\}$$
$$h_{\text{even}} = \{2, 4, 6, ..., 100\}$$

We must explain why we chose $h_{\text{two}} \triangleq$"powers of two", and not, say, $h_{\text{even}} \triangleq$ "even numbers" after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, given that both hypotheses are consistent with the evidence. The key intuition is that we want to avoid **suspicious coincidences**. If the true concept was even numbers, how come we only saw numbers that happened to be powers of two?

assumption, the probability of independently sampling N items (with replacement) from h is given by

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N \tag{3.2}$$

Occam's razor: the simplest solution is always the best

MAP

$$h_{\text{two}} = \{2, 4, 8, 16, 32, 64\}$$
$$h_{\text{even}} = \{2, 4, 6, ..., 100\}$$

Compute likelihoods:

To see how it works, let $\mathcal{D} = \{16\}$. Then $p(\mathcal{D}|h_{\text{two}}) = 1/6$, since there are only 6 powers of two less than 100, but $p(\mathcal{D}|h_{\text{even}}) = 1/50$, since there are 50 even numbers. So the likelihood that $h = h_{\text{two}}$ is higher than if $h = h_{\text{even}}$. After 4 examples, the likelihood of h_{two} is $(1/6)^4 = 7.7 \times 10^{-4}$, whereas the likelihood of h_{even} is $(1/50)^4 = 1.6 \times 10^{-7}$. This is a **likelihood ratio** of almost 5000:1 in favor of h_{two}. This quantifies our earlier intuition that $D = \{16, 8, 2, 64\}$ would be a very suspicious coincidence if generated by h_{even}.

MAP

$$\mathcal{D} = \{16, 8, 2, 64\}$$

$$h_{\text{two}} = \{2, 4, 8, 16, 32, 64\}$$
$$h_{\text{even}} = \{2, 4, 6, ..., 100\}$$

The posterior is simply the likelihood times the prior, normalized. In this context we have

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')} = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(\mathcal{D} \in h')/|h'|^N} \tag{3.3}$$

In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the MAP estimate, i.e.,

$$p(h|\mathcal{D}) \to \delta_{\hat{h}^{MAP}}(h) \tag{3.4}$$

where $\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$ is the posterior mode, and where δ is the **Dirac measure** defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

MAP

In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the MAP estimate, i.e.,

$$p(h|\mathcal{D}) \rightarrow \delta_{\hat{h}^{MAP}}(h) \tag{3.4}$$

where $\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$ is the posterior mode, and where δ is the **Dirac measure** defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

Note that the MAP estimate can be written as

$$\hat{h}^{MAP} = \underset{h}{\text{argmax}} \, p(\mathcal{D}|h)p(h) = \underset{h}{\text{argmax}} \, [\log p(\mathcal{D}|h) + \log p(h)] \tag{3.6}$$

$\log p(\mathcal{D} \mid h)$

depends exponentially on the number of samples N

$p(h)$ independent of N

MAP

In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the MAP estimate, i.e.,

$$p(h|\mathcal{D}) \rightarrow \delta_{\hat{h}^{MAP}}(h) \tag{3.4}$$

where $\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$ is the posterior mode, and where δ is the **Dirac measure** defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

Note that the MAP estimate can be written as

$$\hat{h}^{MAP} = \underset{h}{\text{argmax}} \, p(\mathcal{D}|h)p(h) = \underset{h}{\text{argmax}} \, [\log p(\mathcal{D}|h) + \log p(h)] \tag{3.6}$$

If we have enough data,

MAP estimator $\rightarrow$ MLE estimator : $\underset{h}{\text{argmax}} \log p(\mathcal{D} \mid h)$

data overwhelms the prior !

MAP and MLE

In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the MAP estimate, i.e.,

$$p(h|\mathcal{D}) \rightarrow \delta_{\hat{h}^{MAP}}(h) \tag{3.4}$$

where $\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$ is the posterior mode, and where δ is the **Dirac measure** defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

Note that the MAP estimate can be written as

$$\hat{h}^{MAP} = \underset{h}{\text{argmax}} \, p(\mathcal{D}|h)p(h) = \underset{h}{\text{argmax}} \, [\log p(\mathcal{D}|h) + \log p(h)] \tag{3.6}$$

If we have enough data,

MAP estimator $\rightarrow$ MLE estimator : $\underset{h}{\text{argmax}} \log p(\mathcal{D} \mid h)$

data overwhelms the prior !

Prior and conjugate prior

When the prior and the posterior have the same form, we say that the prior is a **conjugate prior** for the corresponding likelihood. Conjugate priors are widely used because they simplify computation, and are easy to interpret, as we see below.

Suppose $X_i \sim \text{Ber}(\theta)$, where $X_i = 1$ represents "heads", $X_i = 0$ represents "tails", and $\theta \in [0, 1]$ is the rate parameter (probability of heads). If the data are iid, the likelihood has the form

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{3.11}$$

where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails.

prior to use ?

Beta distribution

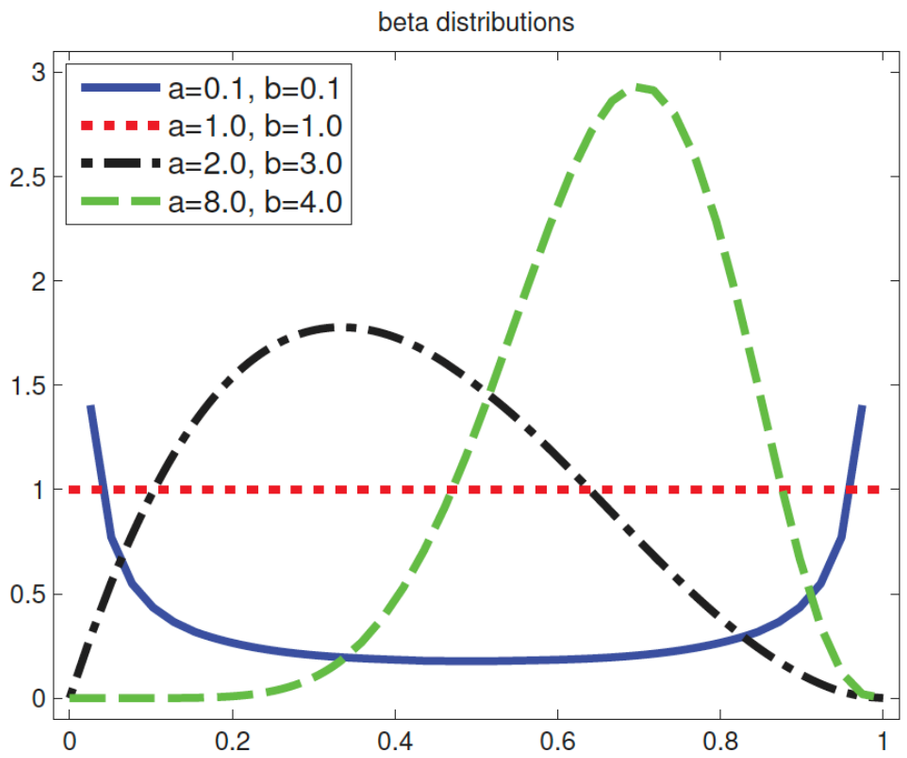
The **beta distribution** has support over the interval $[0, 1]$ and is defined as follows:

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \tag{2.60}$$

Here $B(p, q)$ is the beta function,

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.61}$$

$$\text{mean} = \frac{a}{a+b}, \ \text{mode} = \frac{a-1}{a+b-2}, \ \text{var} = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.62}$$



Beta-Binomial conjugate

Let $X|\theta \sim \text{Binomial}(n, \theta)$ with $\theta \in [0, 1]$. Assume the prior distribution on θ is a Beta distribution:

$$\theta \sim \text{Beta}(\alpha, \beta) \sim \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \ \forall \alpha, \beta > 0,$$

where $\Gamma(\cdot)$ is the Gamma function. Recall

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt = \beta^\alpha \int_0^\infty t^{\alpha-1}e^{-\beta t}dt, \ \forall \alpha, \beta > 0.$$

It is easy to verify that

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \ \forall \alpha > 0,$$

thus Gamma function is usually considered a generalized factorial of positive integers. We also write

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Beta-Binomial conjugate

The mean and variance of θ are given by

$$E(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\alpha} (1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} = \frac{\alpha}{\alpha+\beta},$$

$$\begin{aligned}
Var(\theta) &= E\left(\theta^2\right) - (E(\theta))^2 \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\left(\alpha^2+\alpha\right)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2} \\
&= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.
\end{aligned}$$

Beta-Binomial conjugate

Then we have the joint distribution

$$p(x,\theta) = p(x|\theta)\,p(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

Therefore

$$p(x) = \int_0^1 p(x,\theta)\,d\theta = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}.$$

The posterior distribution

$$p(\theta|x) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \sim \text{Beta}(x+\alpha, n-x+\beta).$$

Beta-Binomial conjugate

A natural estimate of θ given x is its mean

$$\theta_{Bayesian} = E\left(\theta|x\right) = \frac{x+\alpha}{n+\alpha+\beta}.$$

The prior has mean $\alpha/\left(\alpha+\beta\right)$, which is our subjective estimate of θ before we see any data. If we ignore the prior, out estimate of θ would be the sample mean x/n. It turns out Beyesian estimate combines our subjective belief with the information from the data,

$$\theta_{Bayesian} = \frac{n}{n+\alpha+\beta}\cdot\frac{x}{n} + \frac{\alpha+\beta}{n+\alpha+\beta}\cdot\frac{\alpha}{\alpha+\beta},$$

which is a linear combination of the sample mean and prior mean.

Beta-Binomial conjugate

N_1 number of heads, N_0 number of tails

Note that updating the posterior sequentially is equivalent to updating in a single batch. To see this, suppose we have two data sets $\mathcal{D}_a$ and $\mathcal{D}_b$ with sufficient statistics N_1^a, N_0^a and N_1^b, N_0^b. Let $N_1 = N_1^a + N_1^b$ and $N_0 = N_0^a + N_0^b$ be the sufficient statistics of the combined datasets. In batch mode we have

$$p(\theta|\mathcal{D}_a, \mathcal{D}_b) \propto \text{Bin}(N_1|\theta, N_1 + N_0)\text{Beta}(\theta|a, b) \propto \text{Beta}(\theta|N_1 + a, N_0 + b) \tag{3.17}$$

In sequential mode, we have

$$\begin{aligned} p(\theta|\mathcal{D}_a, \mathcal{D}_b) &\propto p(\mathcal{D}_b|\theta)p(\theta|\mathcal{D}_a) && (3.18)\\ &\propto \text{Bin}(N_1^b|\theta, N_1^b + N_0^b)\text{Beta}(\theta|N_1^a + a, N_0^a + b) && (3.19)\\ &\propto \text{Beta}(\theta|\ N_1^a + N_1^b + a, N_0^a + N_0^b + b) && (3.20) \end{aligned}$$

This makes Bayesian inference particularly well-suited to **online learning**, as we will see later.