Deep Learning - Foundations and Concepts

Chapter 2. Probabilities

nonlineark@github

February 4, 2025

Outline

The sum and product rules

- Sum rule: $p(X) = \sum_Y p(X, Y)$.
- Product rule: $p(X, Y) = p(Y|X)p(X)$.

Bayes' theorem

- Bayes' theorem:

$$
\begin{aligned}
p(Y|X) &= \frac{p(X|Y)p(Y)}{p(X)} \\
&= \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}
\end{aligned}
$$

- Prior and posterior probabilities:
    - $p(Y)$ is the prior probability, because it is available *before* we observe the event X.
    - $p(Y|X)$ is the posterior probability, because it is obtained *after* we have observed the event X.

Probability densities

- A probability density $p(x)$ is a real function satisfies the following two conditions[1]:
    - $p(x) \geq 0$.
    - $\int_{-\infty}^{+\infty} p(x) \mathrm{d}x = 1$.
- The *cumulative distribution function* is given by $P(x) = \int_{-\infty}^{x} p(t) \mathrm{d}t$, and usually we have $P'(x) = p(x)$.
- These definitions can easily be extended to higher dimensions.

[1] This is *not* a mathematically robust definition, but it suffices for this course.

Probability densities

- Sum rule: $p(x) = \int p(x, y)\mathrm{d}y$.
- Product rule: $p(x, y) = p(y|x)p(x)$.
- Bayes' theorem: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\mathrm{d}y}$.

Expectations and covariances

- Expectation of f:
  - Discrete case: $E(f) = \sum_x p(x)f(x)$.
  - Continuous case: $E(f) = \int p(x)f(x)\mathrm{d}x$.
- Variance of f: $\mathrm{var}(f) = E((f(x) - E(f))^2) = E(f^2) - E(f)^2$.
- Covariance of:
  - Two random variables: $\mathrm{cov}(x, y) = E((x - E(x))(y - E(y))) = E(xy) - E(x)E(y)$.
  - Two vectors: $\mathrm{cov}(x, y) = E((x - E(x))(y - E(y))^T) = E(xy^T) - E(x)E(y^T)$.

Example distributions

- Uniform distribution: $p(x) = \frac{1}{d-c}, \quad x \in (c, d)$.
- Exponential distribution: $p(x; \lambda) = \lambda \exp(-\lambda x)$.
- Laplace distribution: $p(x; \mu, \gamma) = \frac{1}{2\gamma} \exp(-\frac{|x-\mu|}{\gamma})$.
- Dirac delta function: $p(x; \mu_1, \ldots, \mu_N) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - \mu_n)$.

The Gaussian distribution

- Definition: $\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
- Mean: $E(x) = \int_{-\infty}^{+\infty}\mathcal{N}(x;\mu,\sigma^2)x\mathrm{d}x = \mu$.
- Variance: $\mathrm{var}(x) = E(x^2) - E(x)^2 = \sigma^2$.

Maximum likelihood and its bias

Problem

We have N observations of a random variable x: $x_1, \ldots, x_N$ that are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown. How do we determine these parameters from the data set?

Maximum likelihood and its bias

Problem'

Find μ and σ^2 such that the probability of the data set

$$p(x_1, \ldots, x_N; \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n; \mu, \sigma^2)$$

is maximized.

Maximum likelihood and its bias

Problem"

Let's minimize

$$
\begin{aligned}
L &= -\log p(x_1, \ldots, x_N; \mu, \sigma^2) \\
&= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 + \frac{N}{2} \log \sigma^2 + \frac{N}{2} \log(2\pi)
\end{aligned}
$$

instead.

Maximum likelihood and its bias

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(\mu - x_n) = \frac{N}{\sigma^2}(\mu - \frac{1}{N}\sum_{n=1}^{N} x_n)$$

$$\frac{\partial L}{\partial \sigma} = \frac{N}{\sigma^3}(\sigma^2 - \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2)$$

Setting $\frac{\partial L}{\partial \mu}$ and $\frac{\partial L}{\partial \sigma}$ to 0, we have:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Maximum likelihood and its bias

Let's do some sanity check. Suppose that $x_1, \ldots, x_N$ are generated from a Gaussian distribution whose true parameters are μ and σ^2. We expect the calculated parameters μ_{ML} and σ^2_{ML} to be equal to μ and σ^2 respectively. Or put another way, we expect:

$$E(\mu_{ML}) = \mu$$
$$E(\sigma^2_{ML}) = \sigma^2$$

Is that true?

Maximum likelihood and its bias

$$E(\mu_{ML}) = E(\frac{1}{N}\sum_{n=1}^{N} x_n) = \frac{1}{N}\sum_{n=1}^{N} E(x_n) = \mu$$

$$\begin{aligned} E(\mu_{ML}^2) &= E((\frac{1}{N}\sum_{n=1}^{N} x_n)^2) \\ &= \frac{1}{N^2}(\sum_{1\leq m\neq n\leq N} E(x_m x_n) + \sum_{n=1}^{N} E(x_n^2)) = \mu^2 + \frac{1}{N}\sigma^2 \end{aligned}$$

$$\begin{aligned} E(\sigma_{ML}^2) &= E(\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2) \\ &= \frac{1}{N}\sum_{n=1}^{N} E(x_n^2) - E(\mu_{ML}^2) = \frac{N-1}{N}\sigma^2 \end{aligned}$$

Maximum likelihood and its bias

For a Gaussian distribution, the following estimate for the variance parameter is unbiased:

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma_{ML}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Linear regression from a maximum likelihood perspective

Problem

Assume that given the value of x_n, the corresponding value of t_n has a Gaussian distribution with a mean equal to the value $y(x_n; w)$ and a variance σ^2 (where the parameters w and σ^2 are to be determined). Maximize the likelihood function:

$$p(t, x; w, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(t_n; y(x_n; w), \sigma^2)$$

Linear regression from a maximum likelihood perspective

Again, we minimize the negative log function:

$$
\begin{aligned}
L &= -\log p(t, x; w, \sigma^2) \\
&= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y(x_n; w) - t_n)^2 + \frac{N}{2} \log \sigma^2 + \frac{N}{2} \log(2\pi)
\end{aligned}
$$

We see that maximizing the likelihood function for w is equivalent to minimizing the error function defined by:

$$
E(w) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n; w) - t_n)^2
$$

Linear regression from a maximum likelihood perspective

What has us gained from looking at the linear regression problem from a maximum likelihood perspective? Instead of a point estimate, we now have a predictive distribution:

$$p(\hat{t}, \hat{x}; w_{ML}, \sigma^2_{ML}) = \mathcal{N}(\hat{t}; y(\hat{x}; w_{ML}), \sigma^2_{ML})$$

where

$$\begin{aligned} w_{ML} &= (XX^T)^{-1}Xt \\ \sigma^2_{ML} &= \frac{1}{N}\sum_{n=1}^{N}(y(x_n; w_{ML}) - t_n)^2 \end{aligned}$$

Probability densities are integrand

When changing variable, we need to be aware that probability densities are integrand:

$$p(x)\mathrm{d}x = p(g(y))\mathrm{d}g(y) = p(g(y))g'(y)\mathrm{d}y$$

For multivariate case:

$$p(x)\mathrm{d}x = p(g(y)) \det \frac{\partial(x_1, \ldots, x_N)}{\partial(y_1, \ldots, y_N)} \mathrm{d}y$$

Transformation of densities

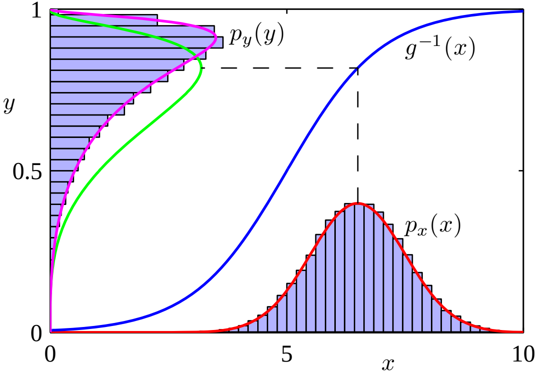
Consider the problem of finding the maximum for a probability density $p(x)$. Say the maximum happens when $x = \hat{x}$. Now we do a change of variable $x = g(y)$, does the maximum for the new probability density happens at $\hat{y}$ where $\hat{x} = g(\hat{y})$?

$$q(y) = p(g(y))g'(y)$$
$$q'(y) = p'(g(y))(g'(y))^2 + p(g(y))g''(y)$$

We see that this is usually not the case, unless g is a linear transformation.

Transformation of densities

Figure: Transformation of the mode of a density



Information

Intuitively, if we have two events x and y that are unrelated, the information gained from observing both of them should be the sum of the information gained from each of them separately:

$$h(x, y) = h(x) + h(y)$$
$$p(x, y) = p(x)p(y)$$

From this it's plausible to define $h(x) = -\log_2 p(x)$.

Entropy

The entropy of a random variable x is defined as the expectation of the information $h(x)$ with respect to the distribution $p(x)$:

$$H[x] = E(h) = \sum_x p(x)h(x) = -\sum_x p(x)\log_2 p(x)$$

When using logarithms to the base of 2, the units of $H[x]$ are bits. From now on, we will switch to the use of natural logarithms in defining entropy, which is measured in units of *nats*.

Maximum entropy for the discrete case

Let $H(p) = -\sum_{n=1}^{N} p_i \log p_i$, where $0 \leq p_i \leq 1$, it's easy to see that $H(p)$ achieves its minimum 0 for unit vectors. When does $H(p)$ achieves its maximum?

Maximum entropy for the discrete case

Finding the maximum of $H(p)$ under the constraint $g(p) = \sum_{n=1}^{N} p_n - 1 = 0$ using Lagrange multiplier:

$$\begin{aligned} \nabla H(p) &= \lambda \nabla g(p) \\ -(\log p_n + 1) &= \lambda \\ p_n &= \frac{1}{N} \\ \max H(p) &= \log N \end{aligned}$$

Differential entropy and its maximum

For the continuous case, we define the differential entropy to be:

$$H[x] = -\int p(x) \log p(x) \mathrm{d}x$$

Differential entropy and its maximum

Finding the maximum of $H(p)$ under the following constraints:

$$\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$$

$$\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x = \mu$$

$$\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\mathrm{d}x = \sigma^2$$

The maximum happens when $p(x)$ is the Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

and

$$\max H(p) = \frac{1}{2}(1 + \log(2\pi\sigma^2))$$

Kullback-Leibler divergence

Problem

Consider some unknown distribution $p(x)$. Suppose we have modelled $p(x)$ using an approximating distribution $q(x)$. If we use $q(x)$ to construct a coding scheme, what is the average additional amount of information required?

Kullback-Leibler divergence

$$
\begin{aligned}
KL(p||q) &= -\int p(x)\log q(x)\mathrm{d}x - (-\int p(x)\log p(x)\mathrm{d}x) \\
&= -\int p(x)\log\frac{q(x)}{p(x)}\mathrm{d}x
\end{aligned}
$$

This is also known as the relative entropy or Kullback-Leibler divergence, or KL divergence, between the distributions $p(x)$ and $q(x)$.

Kullback-Leibler divergence

If f is a convex function, then Jensen's inequality holds:

$$f(E(x)) \leq E(f)$$
$$f(\sum_{n=1}^{N} p_n x_n) \leq \sum_{n=1}^{N} p_n f(x_n)$$
$$f(\int x p(x) \mathrm{d}x) \leq \int p(x) f(x) \mathrm{d}x$$

Notice that $-\log x$ is a convex function, we have:

$$KL(p||q) = \int p(x)(-\log \frac{q(x)}{p(x)}) \mathrm{d}x \geq -\log \int p(x) \frac{q(x)}{p(x)} \mathrm{d}x = 0$$

The equality will hold iff. $q = p$.

Kullback-Leibler divergence

Minimizing the Kullback-Leibler divergence is equivalent to maximizing the likelihood function:

$$KL(p||q) \approx \frac{1}{N} \sum_{n=1}^{N} (-\log q(x_n;\theta) + \log p(x_n))$$

The first term is the negative $\log$ likelihood function for θ under the distribution $q(x;\theta)$ evaluated using the training set.

Conditional entropy

On average, if value for one random variable is already known, what is the additional information needed to specify value for another random variable?

$$
\begin{aligned}
H[y|x] &= -\iint p(x,y)\log p(y|x)\mathrm{d}x\mathrm{d}y \\
H[x,y] &= H[y|x] + H[x]
\end{aligned}
$$

Mutual information

For two random variables, are they "close" to being indepedent?

$$
\begin{aligned}
I[x,y] &= KL(p(x,y)||p(x)p(y)) \\
&= -\iint p(x,y)\log\frac{p(x)p(y)}{p(x,y)}\mathrm{d}x\mathrm{d}y
\end{aligned}
$$

It's easy to see that:

$$
I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]
$$

Model parameters

Denote the training data set by $\mathcal{D}$, and the parameters in the model by w.

- $p(w)$ is our assumptions about w before observing $\mathcal{D}$.
- $p(\mathcal{D}|w)$ is the likelihood function.
- $p(w|\mathcal{D})$ is the uncertainty in w after we have observed $\mathcal{D}$.

We have:

$$
\begin{aligned}
p(w|\mathcal{D}) &= \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})} \\
&= \frac{p(\mathcal{D}|w)p(w)}{\int p(\mathcal{D}|w)p(w)\mathrm{d}w}
\end{aligned}
$$

Regularization

When choosing the model parameters w, instead of maximizing the likelihood function $p(\mathcal{D}|w)$, we maximize the posterior probability $p(w|\mathcal{D})$:

$$-\log p(w|\mathcal{D}) = -\log p(\mathcal{D}|w) - \log p(w) + \log p(\mathcal{D})$$

Say each w_m conforms to a Gaussian distribution:

$$p(w) = p(w;\sigma^2) = \prod_{m=0}^{M} \mathcal{N}(w_m; 0, \sigma^2)$$

Then we have:

$$-\log p(w|\mathcal{D}) = -\log p(\mathcal{D}|w) + \frac{1}{2\sigma^2}\sum_{m=0}^{M} w_m^2 + \text{const}$$

The second term on the right hand side is indeed the penalty term.

Bayesian machine learning

If we are interested in the distribution of t given both x and $\mathcal{D}$, taking into consideration the uncertainty in the value of w, we have the fully Bayesian treatment:

$$p(t|x, \mathcal{D}) = \int p(t|x, w)p(w|\mathcal{D})\mathrm{d}w$$

- The fully Bayesian treatment averages over all possible models:
  - Less likely to lead to over-fitting.
  - Prefer models of intermediate complexity.
- Integrating over the space of parameters is typically infeasible.