Note 12: Gaussian Distribution and Estimations

Math 198: Math for Machine Learning

1 The Gaussian Distribution

So far, we have made reference to the existence of probability distributions, but we have not provided any specific examples (other than occasional allusions to the uniform distribution in homework). There are a litany of distributions which are defined and used in numerous applications of probability, but none is more essential and far-reaching than the *Gaussian (normal) distribution*. This is the familiar "bell curve" distribution.

In the univariate case, the probability distribution function of a Gaussian distribution with mean μ and variance σ^2 is given by

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

The multivariate case for a random vector $\mathbf{X} \in \mathbb{R}^d$ with mean $\mathbf{m} \in \mathbb{R}^d$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is given by

$$p(\mathbf{x};\mathbf{m},\mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma})}}\exp\big(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\big)$$

We write $\mathbf{X} \sim \mathcal{N}(\mathbf{m},\mathbf{\Sigma})$ to signify that $\mathbf{X}$ is normally distributed with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$.

Consider the univariate case. We can define $\bar{x} = x - \mu$ and

$$g(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\big(\frac{-z}{2}\big)$$

Then

$$p(x;\mu,\sigma^2) = g\big(\frac{\bar{x}^2}{\sigma^2}\big)$$

Note that g is strictly monotonically decreasing in its argument. So p will give us smaller probabilities for points x further from the mean μ, and the rate at which these probabilities will decrease is determined by σ^2. We can extend this to the univariate case by defining $\bar{\mathbf{x}} = \mathbf{x} - \mathbf{m}$ and

$$g(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma})}}\exp\big(-\frac{\mathbf{z}}{2}\big)$$

Then

$$p(\mathbf{x};\mathbf{m},\mathbf{\Sigma}) = g(\bar{\mathbf{x}}^\top \mathbf{\Sigma}^{-1}\bar{\mathbf{x}})$$

Once again, g is strictly monotonically decreasing in its argument. Smaller values of $\bar{\mathbf{x}}^\top \mathbf{\Sigma}^{-1}\bar{\mathbf{x}}$, and thus higher probabilities, correspond to values of the random vector $\mathbf{X}$ which are closer to its mean, with the rate again being determined by the covariance matrix.

2 Estimation

2.1 Maximum Likelihood Estimation

A frequent use case of probability theory is fitting models to data. We may have an idea of which model to use to predict future values of random variables, but don't know the specific values of the parameters used

to specify the model. We can then estimate values of these parameters based on the data we've already seen, using various methods.

One method of parameter estimation is to select the parameters which maximize the probability of observing the data we have observed. Suppose we have observed the values $x_1, \ldots, x_n$ of independent, identically distributed random variables $X_1, \ldots, X_n$. We define the *likelihood function*

$$\mathcal{L}(\theta) = p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

with the parameters denoted by θ; the maximum likelihood parameters are thus $\bar{\theta}_{\text{MLE}} = \arg\max_\theta \mathcal{L}(\theta)$.

It is often simpler to maximize the log-likelihood

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta)$$

The optimal $\bar{\theta}_{\text{MLE}}$ will be the same in either case, as probabilities are always positive, and the logarithm is a monotonically increasing function. For some models, maximizing $\log \mathcal{L}(\theta)$ can be done analytically, as $\mathcal{L}$ is differentiable; for others, numerical approaches are necessary.

2.2 Maximum A Posteriori Estimation

If we have a notion of which values of the parameters are more likely than others, then we can use Bayes' rule to compute the posterior distribution of the parameters given the observed data:

$$p(\theta | x_1, \ldots, x_n) = \frac{p(\theta) p(x_1, \ldots, x_n | \theta)}{p(x_1, \ldots, x_n)}$$

Because the denominator of this term does not depend on the value of θ, we can just maximize the numerator with respect to θ, and ignore the constant:

$$\bar{\theta}_{\text{MAP}} = \arg\max_\theta p(\theta) p(x_1, \ldots, x_n | \theta)$$

Assuming the observations are i.i.d, we can rewrite using log-likelihoods into the more tractable form

$$\bar{\theta}_{\text{MAP}} = \arg\max_\theta \left(\log p(\theta) + \sum_{i=1}^{n} \log p(x_i | \theta) \right)$$

Applications: OLS, One Last Time

Way back at the start of the course, we attempted to predict a basketball player's plus-minus stat given their points, assists, and rebounds. Recall the following passage from note 1:

> Observe that p(data = $\mathbf{X}, \mathbf{y}$ | weights = $\mathbf{w}$) represents the probability that we observed our data given our weights. We seek the weights which maximize this probability, as these will generalize the best to new data. That is, we wish to find $\hat{\mathbf{w}}$ which satisfies
>
> $$\hat{\mathbf{w}} = \max_{\mathbf{w}} p(\text{data} = \mathbf{X}, \mathbf{y} \mid \text{weights} = \mathbf{w})$$
>
> We will later prove that this is equivalent to finding the weights which minimize the sum of the squared differences between our predictions and our observations:
>
> $$\hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$$

Well, it's finally time to complete this proof! We can now recognize the method by which we can approach this argument, by framing this as a maximum likelihood estimation problem. Our statistical model assumes the labels and features are related as

$$y_i = \mathbf{x_i}^\top \mathbf{w} + \epsilon_i$$

where ϵ_i is a random, normally distributed noise variable; that is, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some σ^2. What, then, is our probability distribution for some observation y_i? The probability that we observe y_i given $\mathbf{x_i}$ and $\mathbf{w}$ is exactly the probability that ϵ_i takes on the value $y_i - \mathbf{x_i}^\top \mathbf{w}$. So we can write

$$p(y_i | \mathbf{x_i}; \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mathbf{x_i}^\top \mathbf{w})^2}{2\sigma^2})$$

We can then maximize the log-likelihood

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, \sigma^2) &= \sum_{i=1}^{n} \log p(y_i | \mathbf{x_i}; \mathbf{w}, \sigma^2) \\
&= \sum_{i=1}^{n} \log(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \mathbf{x_i}^\top \mathbf{w})^2}{2\sigma^2})) \\
&= -(\frac{n}{2} \log 2\pi + n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x_i}^\top \mathbf{w})^2)
\end{aligned}$$

Note that, since σ and n are constants, this is equivalent to minimizing $\sum_{i=1}^{n} (y_i - \mathbf{x_i}^\top \mathbf{w})^2 = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$. Using either calculus or the linear algebra techniques we used earlier in the course, we can then derive that the optimal weights $\hat{\mathbf{w}}$ are given by

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$