C1-8

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

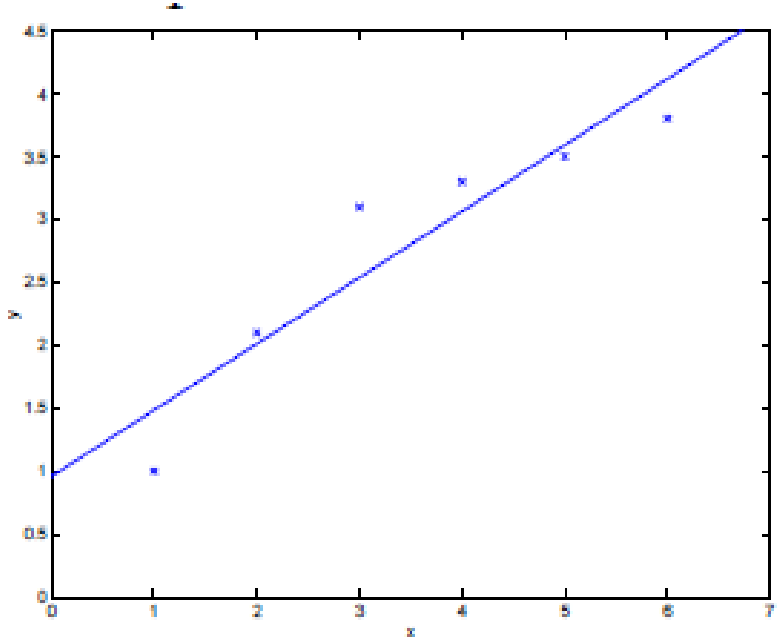
2016.9.1

第九集：经验风险最小化

- 偏差与方差的权衡(Bias/Variance Tradeoff)
- 联合界与Hoeffding不等式(Union Bound/Hoeffding Inequality)
- 经验风险最小化(Empirical Risk Minimization (ERM))
- 一致收敛(Uniform Convergence)

第九集：经验风险最小化

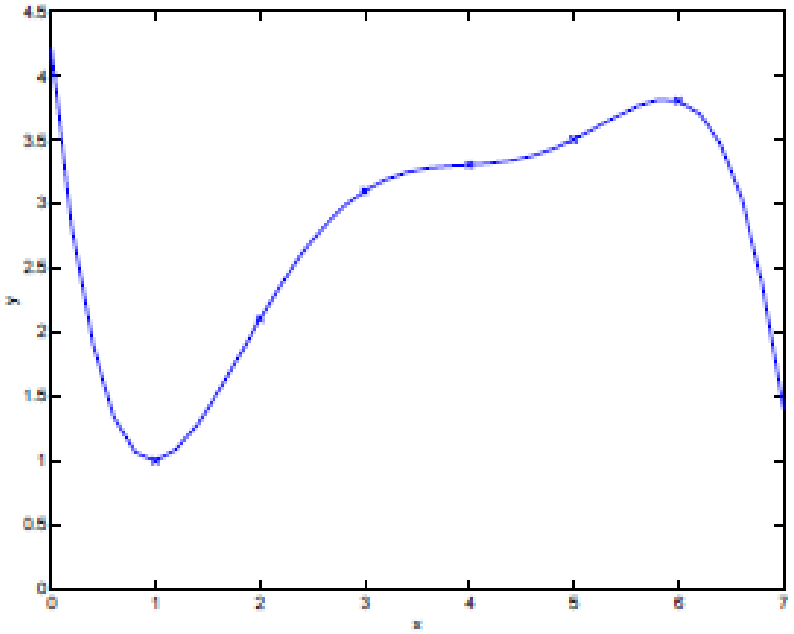
- 偏差与方差的权衡
  - 接下来的内容将从学习理论(learning theorem)的角度对于机器学习的众多问题进行分析
  - 首先先来看欠拟合和过拟合的问题



左图中试图用一阶直线拟合二次函数分布的样本，因此即使我们用大量的（甚至无穷多的）样本拟合出一条一阶直线依然不能准确反映数据的特征，出现了欠拟合问题。我们对于这样的用无穷多个样本进行拟合仍然出现较大错误的问题定义为这种拟合的偏差(bias)很大（因此偏差对应的是欠拟合问题）

第九集：经验风险最小化

- 偏差与方差的权衡
  - 首先先来看欠拟合和过拟合的问题



左图中试图用高阶曲线（图中为五阶曲线）拟合二次函数分布的样本，因此对于这个小的训练集上的样本能够做到完美拟合，但是对于其他数据样本来说仍不能正确反映分布的情况（因为这些数据样本很可能也不准确，带有随机噪声），出现了过拟合问题。我们对于这样的用小的训练集上的样本能够做到完美拟合，但是对于其他数据样本来说仍不能正确反映分布的情况定义为这种拟合的方差(variance)很大（因此方差对应的是过拟合问题）

第九集：经验风险最小化

- 偏差与方差的权衡
  - 由于欠拟合和过拟合都不是我们想要的结果，因此考虑对于偏差和方差进行统一考量，引入泛化误差(generalization error)的概念
    - 泛化误差是一个描述学生机器在从样品数据中学习之后，离教师机器之间的差距的函数。简单理解就是这个建模和标准建模之间的差距
    - 因此我们对于一个机器学习方法来说，对于泛化误差进行优化相当于对于偏差与方差进行权衡，从而找到最优拟合方式。而这个问题又可以分解为下面三个子问题：
      - 1、如何对于泛化误差进行建模从而正确体现偏差与方差的权衡
      - 2、对于整个机器学习问题来说我们希望泛化误差尽量小，但是在实际操作时我们都是基于训练集根据训练的误差进行的训练和模型选择，那怎么把训练误差和泛化误差有机结合在一起
      - 3、如何使用泛化误差对于某一个机器学习方法进行评价

第九集：经验风险最小化

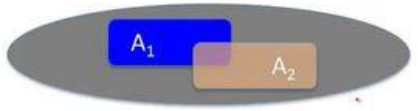
- 联合界与Hoeffding不等式
  - 在开始探讨上面三个子问题前，先来看两个基础的引理
    - 联合界引理：假设有k个不同的事件{A1, A2, ..., Ak}（事件相互之间不一定独立），那么这k个事件同时发生的概率小于等于这k个事件分别发生的概率和（等于在事件相互独立时取得）

$$P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \ldots + P(A_k)$$

The union bound

- For events A_1 and A_2

$$\Pr[A_1 \cup A_2] \leq \Pr[A_1] + \Pr[A_2]$$



第九集：经验风险最小化

- 联合界与Hoeffding不等式
  - 而在介绍下一个引理Hoeffding不等式之前先介绍两个基本的引理：Marcov不等式和Chebyshev不等式
    - Marcov不等式：令X为非负随机变量，则对任意t>0，有：

$$p(X \geq t) \leq \frac{\mathrm{E(X)}}{t}$$

    - 证明：

由于随机变量 X 非负，因此对于 E(X)的积分区域计算只需要考虑 x ≥0 的部分

$$\mathrm{E(X)} = \int_0^\infty xp(x)dx$$

由于 t>0，因此可将积分区域分为正的两项相加（因为 x 和 p(x)均非负）

$$\int_0^\infty xp(x)dx = \int_0^t xp(x)dx + \int_t^\infty xp(x)dx \geq \int_t^\infty xp(x)dx$$

由于此时 x 的积分区间为[t,∞)，即 x≥t，因此有：

$$\int_t^\infty xp(x)dx \geq t\int_t^\infty p(x)dx = tp(X \geq t)$$

因此：

$$p(X \geq t) \leq \frac{\mathrm{E(X)}}{t}$$

第九集：经验风险最小化

- 联合界与Hoeffding不等式
  - Marcov不等式
    - 从Marcov不等式中不难发现对于非负随机变量X来说P(X≥E(X))≤1，而且随着X的进一步增大其概率的上界将不断缩小
    - 与此同时还可得出此时的X不止可以是非负随机变量，还可以是一个非负函数，而由此还引出了Chebyshev不等式
      - Chebyshev不等式：对于随机变量X和任意t>0，有：

$$p\left(\left|X-\mu\right|\geq t\right)\leq\frac{\sigma^2}{t^2}$$

      - 证明：

其中μ和σ分别为变量X的均值与方差

$$p\left(\left|X-\mu\right|\geq t\right)=p\left(\left(X-\mu\right)^2\geq t^2\right)\leq\frac{\mathrm{E}\left(\left(X-\mu\right)^2\right)}{t^2}=\frac{\sigma^2}{t^2}$$

第九集：经验风险最小化

- 联合界与Hoeffding不等式
  - 对于Chebyshev不等式来说给出了随机变量X的分布的概率上界，但这个界比较松，而在Hoeffding不等式中则给出了一个更紧的概率上界
  - 而Hoeffding不等式基于Hoeffding引理，其表述如下：
    - Hoeffding引理：对于均值为0（E(X)=0）的随机变量X，并保证X的取值范围在[a, b]之间，则对于任意的λ有：

$$\mathrm{E}\left[e^{\lambda x}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

由于$\mathrm{f}(\mathrm{x}) = e^{\lambda x}$为凸函数，因此对于任意定义域上的 x 和$0 \leq \theta = \frac{b-x}{b-a} \leq 1$有：

$$e^{\lambda x} = \mathrm{f}(\theta \mathrm{a} + (1-\theta)\mathrm{b}) \leq \theta \mathrm{f}(\mathrm{a}) + (1-\theta)\mathrm{f}(\mathrm{b}) = \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$$

由于$\mathrm{a} \leq \mathrm{x} \leq \mathrm{b}$恒成立，因此$\mathrm{a} \leq \mathrm{E}(\mathrm{x}) \leq \mathrm{b}$，因此有：

$$E\left(e^{\lambda x}\right) = e^{\lambda E(x)} \leq \frac{b-E(x)}{b-a}e^{\lambda a} + \frac{E(x)-a}{b-a}e^{\lambda b}$$

由于$E(x) = 0$因此

$$E\left(e^{\lambda x}\right) \leq \frac{b}{b-a}e^{\lambda a} + \frac{a}{b-a}e^{\lambda b} = \left(-\frac{a}{b-a}\right)e^{\lambda a}\left(-\frac{b}{a} + e^{\lambda b - \lambda a}\right) = \left(-\frac{a}{b-a}\right)e^{\lambda a}\left(-\frac{b-a+a}{a} + e^{\lambda b - \lambda a}\right)$$

$$= \left(-\frac{a}{b-a}\right)\left(-\frac{b-a}{a} - 1 + e^{\lambda(b-a)}\right)e^{\lambda a} = (1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)})e^{\lambda a}$$

此时令$\theta = -\frac{a}{b-a}$有：

$$E\left(e^{\lambda x}\right) \leq (1-\theta+\theta e^{\lambda(b-a)})e^{-\lambda\theta(\mathrm{b}-a)}$$

由于右式大于 0，因此$\exp\left[\log\left(\left(1-\theta+\theta e^{\lambda(b-a)}\right)e^{-\lambda\theta(\mathrm{b}-a)}\right)\right] = (1-\theta+\theta e^{\lambda(b-a)})e^{-\lambda\theta(\mathrm{b}-a)}$，而$\log\left(\left(1-\theta+\theta e^{\lambda(b-a)}\right)e^{-\lambda\theta(\mathrm{b}-a)}\right) = -\lambda\theta\left(\mathrm{b}-a\right) + \log\left(1-\theta+\theta e^{\lambda(b-a)}\right)$，令$\mathrm{u} = \lambda\left(b-a\right)$则可将其表示为$\mathrm{h}(\mathrm{u}) = -\theta\mathrm{u} + \log\left(1-\theta+\theta e^{u}\right)$，根据带拉格朗日余项的 Taylor 展开有：

$$\exists \mathrm{v}, 0 \leq \mathrm{v} \leq \mathrm{u}, \mathrm{h}(\mathrm{u}) = \mathrm{h}(0) + \mathrm{uh}'(0) + \frac{1}{2}u^2h''(v) = \frac{1}{2}u^2\frac{\theta e^v}{1-\theta+\theta e^v}\left(1-\frac{\theta e^v}{1-\theta+\theta e^v}\right) = \frac{1}{2}u^2t(1-t) \leq \frac{1}{8}u^2$$

$$= \frac{1}{8}\lambda^2(b-a)^2$$

由此可得：

$$\mathrm{E}\left[e^{\lambda x}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

第九集：经验风险最小化

- 联合界与Hoeffding不等式
  - Hoeffding不等式：对于服从某个分布的随机变量X，其经验期望和真实期望满足下面不等式关系

随机变量 X 的经验期望$\overline{X}$为抽取得到一组独立的样本{x1, ⋯, xn}的均值，即：

$$\overline{X}=\frac{\sum_{i=1}^{n} xi}{n}$$

而经验期望与分布的真实期望满足（其中ai $\leq xi \leq bi$，t >0）：

$$p\left(\left|\overline{X}-E\left(\overline{X}\right)\right| \geq \mathrm{t}\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(bi-ai)^2}\right)$$

$$p\left(\left|\overline{X}-E\left(\overline{X}\right)\right| \geq \mathrm{t}\right)=p\left(\overline{X}-E\left(\overline{X}\right) \geq \mathrm{t}\right)+p\left(E\left(\overline{X}\right)-\overline{X} \geq \mathrm{t}\right)$$

其中由于 $\overline{X}=\frac{\sum_{i=1}^{n} xi}{n}$，因此$\overline{X}-E\left(\overline{X}\right)=\frac{S_n-E(S_n)}{n}$，其中$S_n=\mathrm{x1}+\cdots+\mathrm{xn}$，引入$\lambda>0$有：

$$p\left(\overline{X}-E\left(\overline{X}\right) \geq \mathrm{t}\right)=p\left(e^{\lambda\left(\overline{X}-E\left(\overline{X}\right)\right)} \geq e^{\lambda t}\right)=p\left(e^{\lambda\left(S_n-E(S_n)\right)} \geq e^{\lambda tn}\right)$$

根据 Marcov 不等式有：

$$p\left(e^{\lambda\left(S_n-E(S_n)\right)} \geq e^{\lambda tn}\right) \leq \frac{E\left(e^{\lambda\left(S_n-E(S_n)\right)}\right)}{e^{\lambda tn}}$$

又根据抽取样本的独立性有：

$$\frac{E\left(e^{\lambda\left(S_n-E(S_n)\right)}\right)}{e^{\lambda tn}}=\frac{\prod_{i=1}^{n} E\left(e^{\lambda(\mathrm{xi}-E(xi))}\right)}{e^{\lambda tn}}$$

而分子中由于$\mathrm{E}\left(\mathrm{xi}-E(xi)\right)=0$因此可以使用 Hoeffding 引理进行计算：

$$\frac{\prod_{i=1}^{n} E\left(e^{\lambda(\mathrm{xi}-E(xi))}\right)}{e^{\lambda tn}} \leq e^{-\lambda \mathrm{tn}} \prod_{i=1}^{n} \exp\left(\frac{\lambda^2(bi-ai)^2}{8}\right)$$

同理可得：

$$p\left(E\left(\overline{X}\right)-\overline{X} \geq \mathrm{t}\right) \leq e^{-\lambda \mathrm{tn}} \prod_{i=1}^{n} \exp\left(\frac{\lambda^2(bi-ai)^2}{8}\right)$$

因此：

$$p\left(\left|\overline{X}-E\left(\overline{X}\right)\right| \geq \mathrm{t}\right) \leq 2e^{-\lambda \mathrm{tn}} \prod_{i=1}^{n} \exp\left(\frac{\lambda^2(bi-ai)^2}{8}\right)=2\exp\left(-\lambda \mathrm{tn}+\sum_{i=1}^{n} \frac{\lambda^2(bi-ai)^2}{8}\right)$$

$$\leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(bi-ai)^2}\right)$$

第九集：经验风险最小化

- 联合界与Hoeffding不等式
  - 通过Hoeffding不等式不难得到如下的结论，假设Z1,...,Zm为从某一二项分布Bernoulli(φ)进行m次独立抽取的结果，那么这m个样本的经验期望与分布的真实期望φ之间满足下面不等式关系（其中γ>0）

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

    - 从中不难发现，随着抽样次数的增多，样本的经验期望与分布的真实期望之间的差值限制在很小的范围内的概率将不断增大
    - 这个结论在学习理论中也被称为Chernoff界(Chernoff bound)

第九集：经验风险最小化

- 经验风险最小化
  - 接下来将对于一个具体机器学习的方法进行分析
    - 假设一个二元分类问题(y∈{0,1})和m个样本的训练集S，可以定义训练误差(training error，也被称为经验风险(empirical risk)、经验误差(empirical error))为错误分类样本占全部样本的比例

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$$

    - 但这只是基于训练集的误差，实际上我们更关心的是对于全部数据集的误差，因此可以定义泛化误差为从某个分布D的数据集中任意抽取一个新的样本划分错误的概率（不论偏差的增大还是方差的增大都会导致泛化误差的增大）

$$\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

      - 在这里面要求进行训练的数据集和进行测试的数据集是基于同个分布的数据集（PAC假设中有要求）

第九集：经验风险最小化

- 经验风险最小化
  - 但是对于一个具体的学习问题来说泛化误差往往难以衡量，因此在训练过程中都用训练误差进行衡量

$$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta)$$

    - 我们称此过程为经验风险最小化(ERM, empirical risk minimization)
    - 更进一步，对于一个具体的学习问题来说我们不仅需要训练参数。更重要地，我们还需要选择一个合适的模型进行训练。因此将这些因素进行统一考虑，提出假设类(hypothesis class)这一概念
      - 假设类H为一系列假设的集合，每一种假设都对应了一个特定的模型和训练参数，比如对于二元分类问题来说，假设类H可定义为

$$\mathcal{H} = \{h_\theta : h_\theta(x) = 1\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{n+1}\}$$

      - 而我们的目标在于在假设类H中找到能使得经验风险最小化的假设

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

第九集：经验风险最小化

- 经验风险最小化
  - 首先先考虑假设类中假设有限的情况，目标在于从这有限个假设中选出能使经验风险最小化的那个假设
    - 但是对于一个机器学习模型来说泛化误差比经验误差更能准确地反映模型的准确度，那么就需要探讨下面两个问题
      - 1、经验风险（训练误差）和泛化误差之间有什么关系
      - 2、泛化误差的界是什么
    - 对于某一个假设hi而言，我们可以将任意训练集中样本的分类正确或者错误看成是一个二项分布的随机事件，比如记第j次独立抽取训练样本分类错误的事件记为Zj(Zj=1{hi(xj)≠yj})
      - 由此可以将训练误差表达为：

$$\hat{\varepsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$$

第九集：经验风险最小化

- 经验风险最小化
  - 根据Hoeffding不等式可以得到

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

    - 这表明对于某个假设hi来说，如果训练样本集合足够大，那么训练误差接近泛化误差的概率非常大，但是我们更加关心的是对于一个假设类训练误差和泛化误差之间的关系
    - 因此我们用随机变量Ai表示在假设hi下训练误差和泛化误差距离大于γ的事件，即P(Ai)≤2exp(-2γ^{2}m).由此可以借助于联合界引理衡量存在某个假设，其训练误差和泛化误差距离大于γ的概率

$$\begin{aligned} P(\exists h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \\ &= 2k\exp(-2\gamma^2 m) \end{aligned}$$

第九集：经验风险最小化

- 一致收敛
  - 通过衡量假设类中存在某个假设，其训练误差和泛化误差距离大于γ的概率，可以进一步得出假设类中所有假设的训练误差和泛化误差的距离小于等于γ的概率

$$\begin{aligned} P(\neg\exists\, h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\ &\geq 1 - 2k\exp(-2\gamma^2 m) \end{aligned}$$

    - “假设类中所有假设的训练误差和泛化误差的距离小于等于γ”这一结果也被称为一致收敛(uniform convergence)
    - 从中也可以看出在给定训练误差和泛化误差的距离γ，要求假设类中所有假设的训练误差和泛化误差的距离小于等于γ的概率为1-δ的情况下至少需要的样本数应满足

$$\delta = 2k\exp(-2\gamma^2 m) \longrightarrow m \geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta}$$

      - 这式子给出了给问题所需的训练样本数，而这也被称为算法的采样复杂度(sample complexity)

第九集：经验风险最小化

- 一致收敛
  - 除了可以得到样本复杂度外，同样可以估计给定样本数目m，要求假设类中所有假设的训练误差和泛化误差的距离小于等于γ的概率为1-δ的情况下对应的训练误差和泛化误差的距离γ至少为：

$$\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}.$$

  - 由此我们得到了训练误差和泛化误差的关系，但由于训练时我们使用的仍然为训练误差，那么在一致收敛成立的情况下训练误差最优的算法是否代表着泛化误差最优？它们之间有什么关联？

第九集：经验风险最小化

- 一致收敛
  - 此时我们假设h^为假设集中训练误差最优的假设，h*为假设集中泛化误差最优的假设，对于这二者存在下面的不等式关系

$$\begin{aligned}\varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma\end{aligned}$$

这里面考虑的是最坏情况，其中第一个不等式来自于一致收敛情况下所有假设的训练误差和泛化误差的距离小于等于γ；第二个不等式由于h^为假设集中训练误差最优的假设，因此对应的训练误差值应该是所有假设中最小的；第三个不等式同样来自于一致收敛，所有假设的训练误差和泛化误差的距离小于等于γ

  - 从以上不等式关系中可以看出在最坏情况下用训练误差得到的最优假设与泛化误差最优的假设的泛化误差间距离小于等于2γ
    - 从而可得定理：在给定样本数目m，要求假设类中所有假设的训练误差和泛化误差的距离小于等于γ的概率为1-δ的情况下满足下面的不等式关系

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

从这里也能看出偏差和方差的权衡关系，当我们选取一个更大的假设集时，右式的左半项只减不增（因为从更大的范围取最小值），右半项只增不减（因为假设数目k在分子），而这左半项与偏差有关，右半项与方差有关

第九集：经验风险最小化

- 一致收敛
  - 同样地，对于训练样本数也有类似结论
    - 在给定训练误差和泛化误差的距离γ，要求训练误差得到的最优假设与泛化误差最优的假设的泛化误差间距离小于等于2γ的概率至少为1-δ的情况下至少需要的样本数应满足：

$$
\begin{aligned}
m &\geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta} \\
&= O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right)
\end{aligned}
$$